From Raw Data to Physics Results (1/3)

Paul Laycock

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Credits for material

- **All ATLAS, CMS and NA62 material is copyright CERN**
  - I invite you to look for more on the CERN site, it's a very useful resource

- **Belle II material taken from Belle II public web pages**

- I have relied heavily on material from
  - **Anna Sfyrla** and **Jamie Boyd** from previous SSLP courses
  - Thanks also to **Dave Barney**

- I do not always credit original sources when the provenance was lost
  - apologies where that is the case

Assumptions and disclaimers

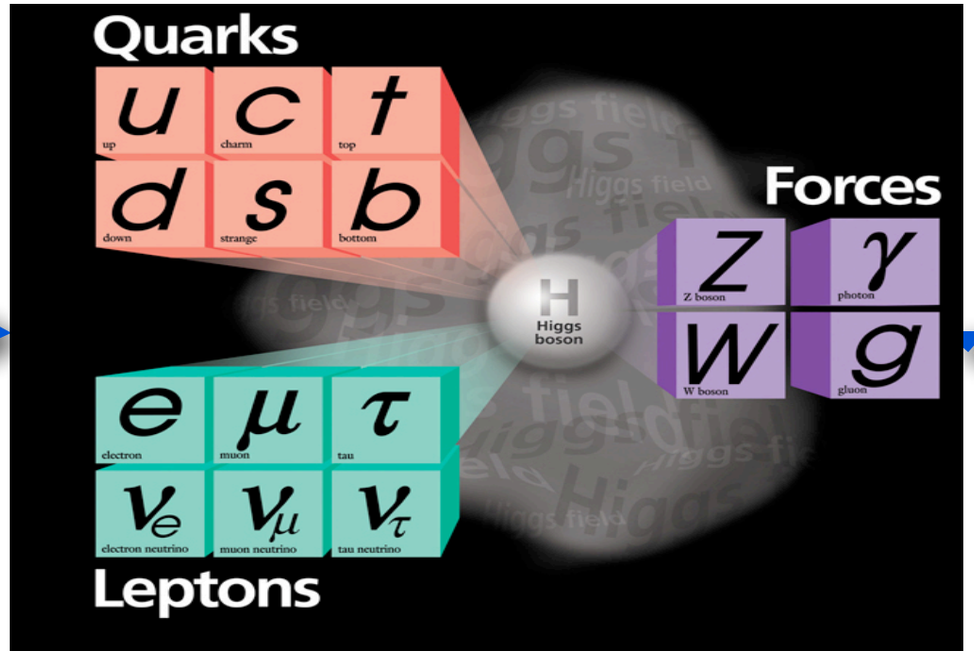
- **Aims**
  - Learn about the journey of data, from the raw data from the detectors that make up our experiments, to the highly refined data we publish in scientific journals

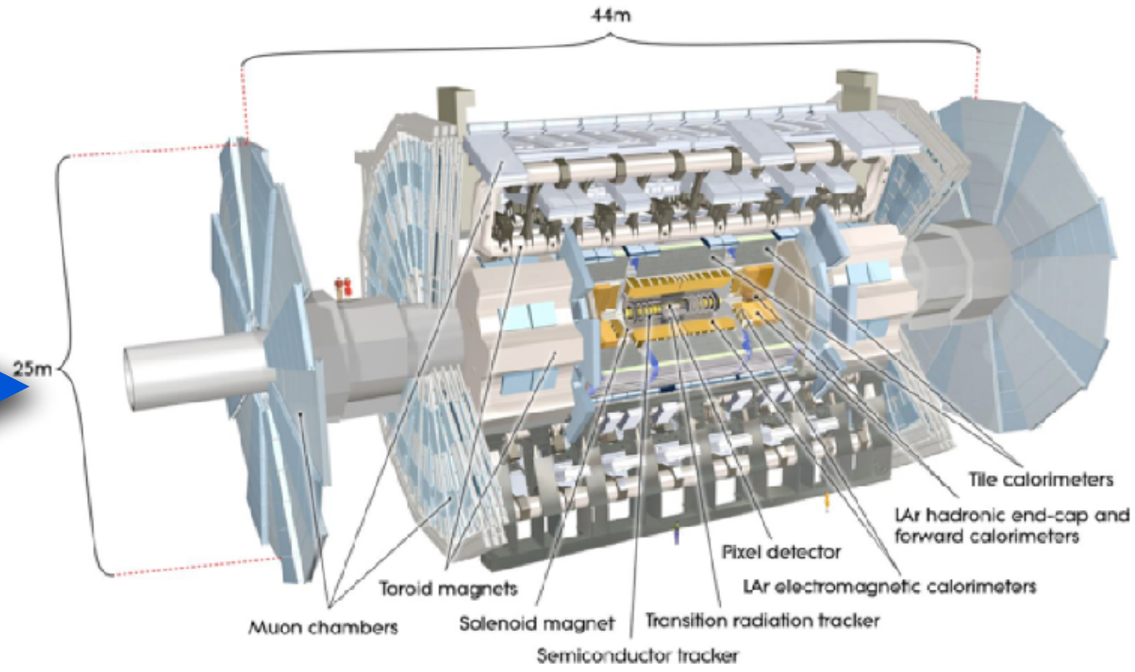
- **Assumptions**
  - You have some (first) ideas about particle physics and the questions that we're trying to answer at **CERN** and worldwide
  - You know something about high energy physics detectors and their data

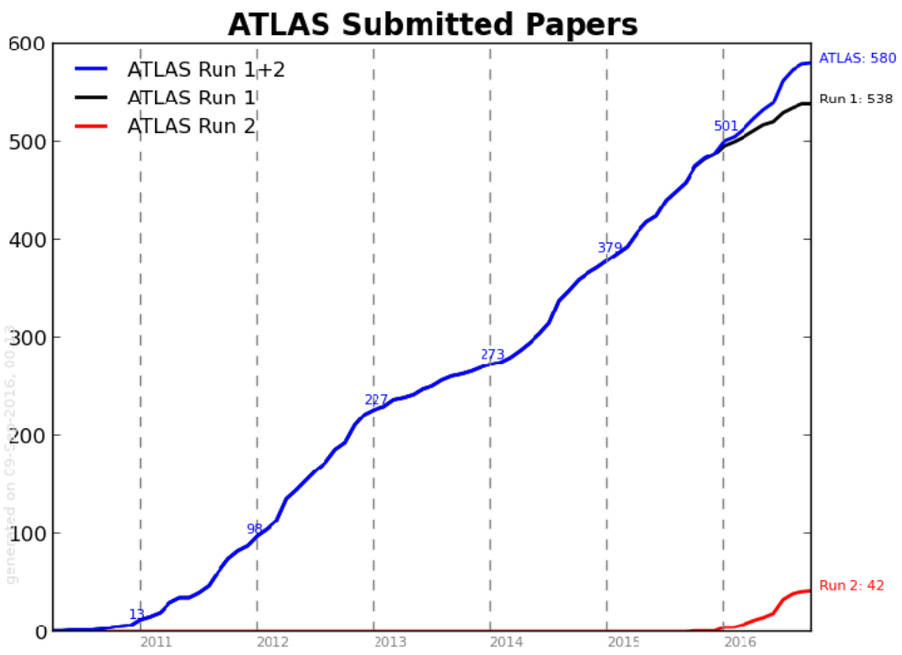
- **Disclaimer**
  - Choice of examples is often based on those experiments I've personally worked on (**H1, ATLAS, NA62** and **Belle II**)

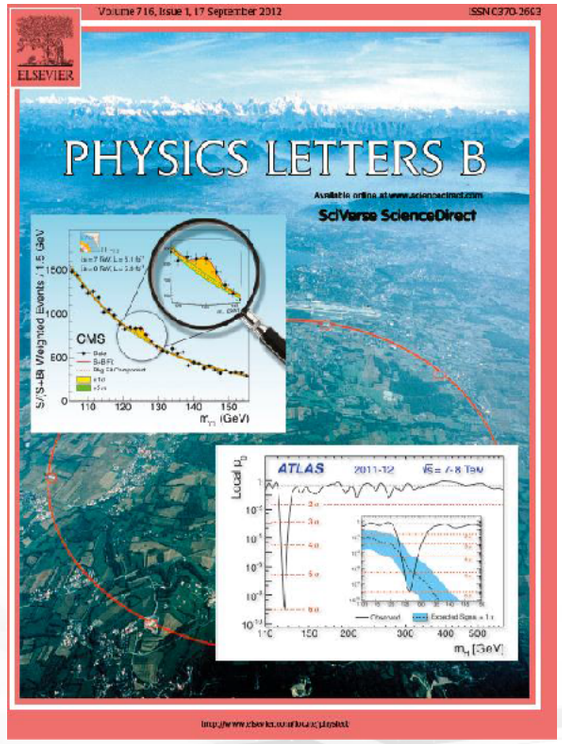
- Feedback and questions are welcome: paul.james.laycock@cern.ch
  - *I will be here all week, email is a good way to contact me*

The physics cycle





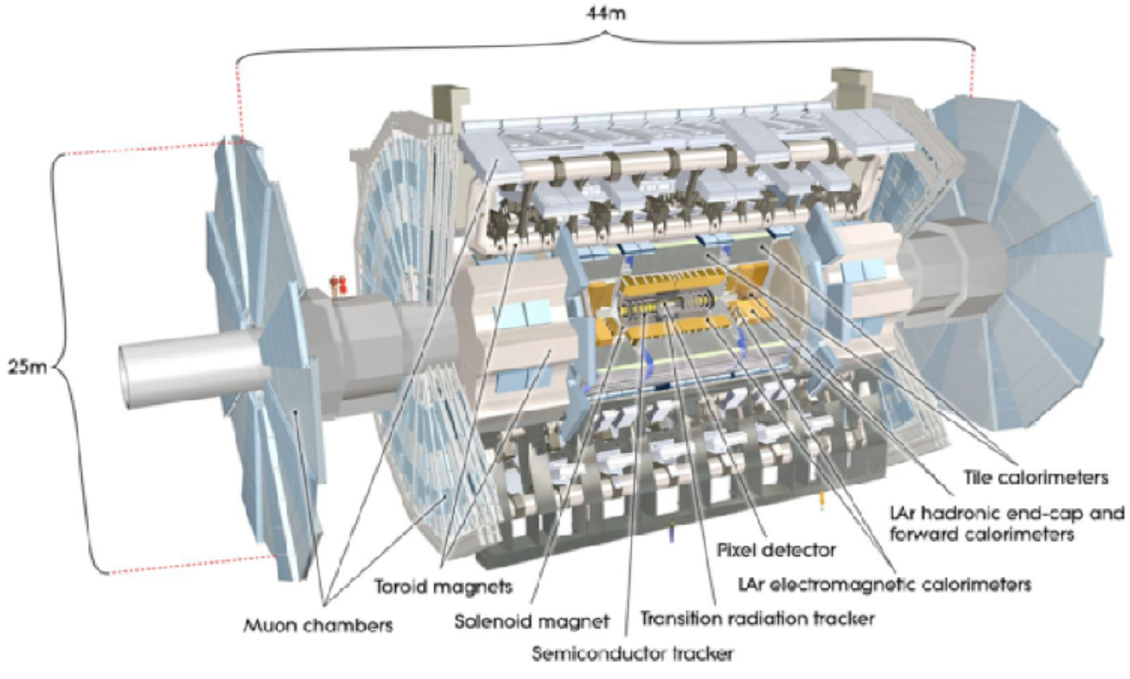




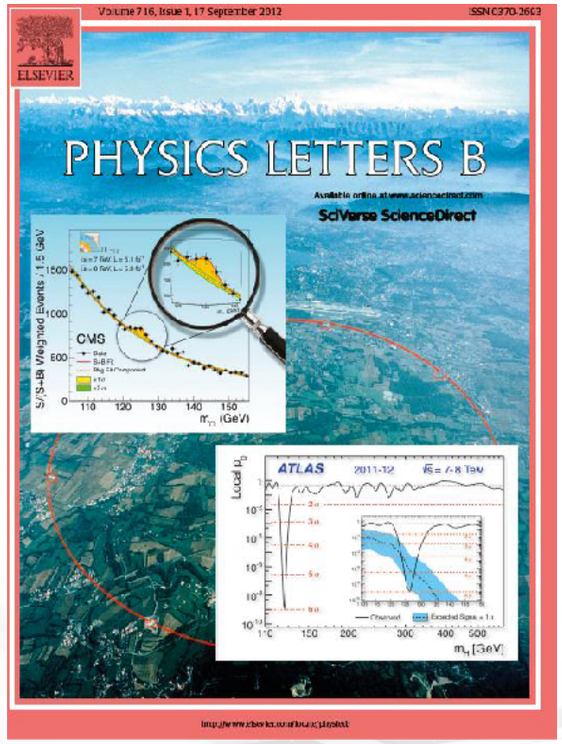
Experimental physics

- Much of the work of the experimental physicist is running experiments and extracting measurements from them

- ***Note*** *- Experimental physicists also need to propose, design and build new experiments (see previous slide)*



- These lectures are focused on understanding how we turn raw experimental detector data into physics results that we can publish
  - Results must be ***accurate***
  - with well understood ***precision***
  - It's important to understand the difference between these two words, we often confuse them

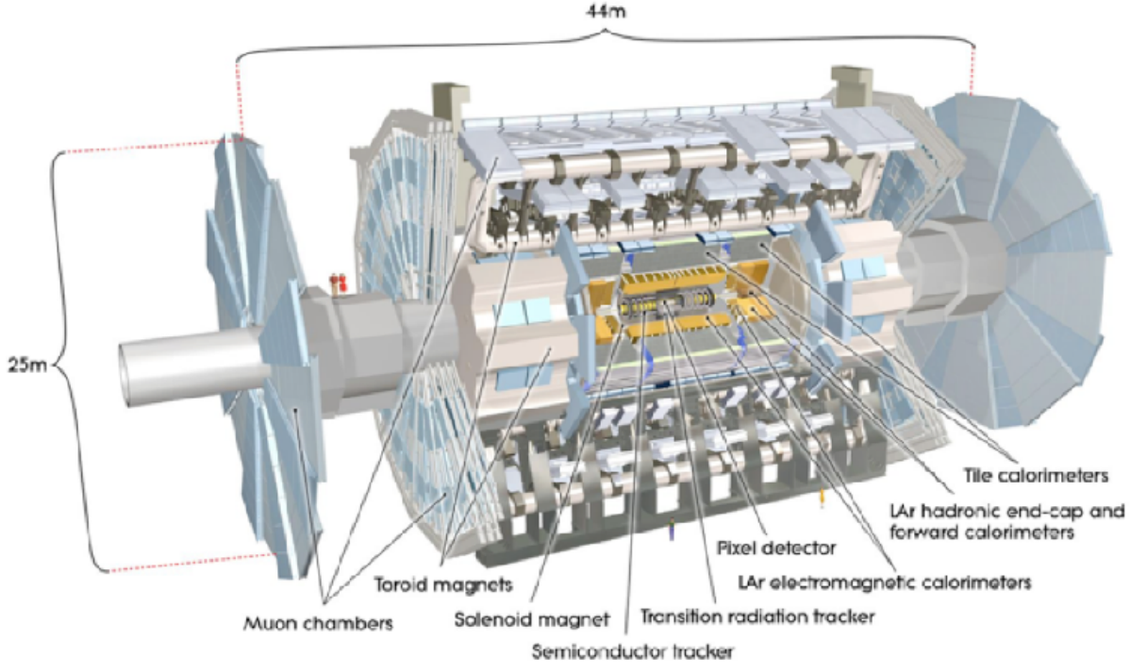


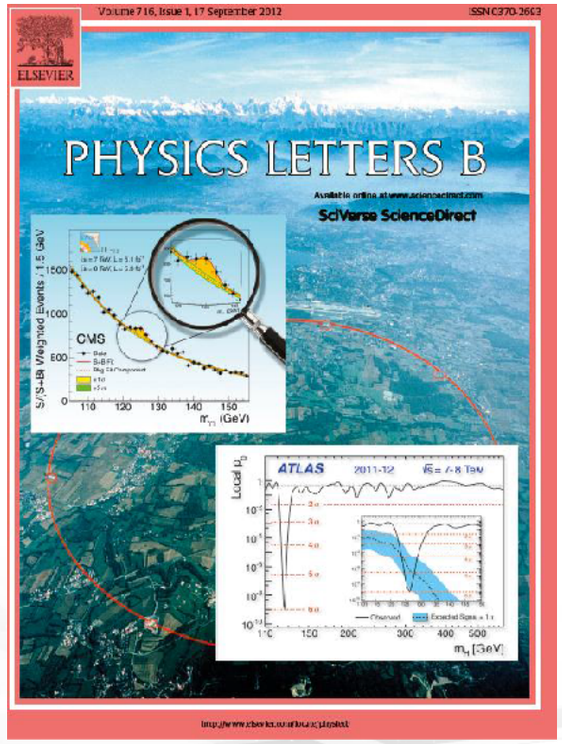
Course outline

- **Lecture 1**
  - The journey of raw data from the detector to a publication

- **Lecture 2**
  - How we reconstruct fundamental physics processes from raw detector data

- **Lecture 3**
  - How we extract our signals from the mountain of data, finding needles in the haystack





Testing theoretical predictions

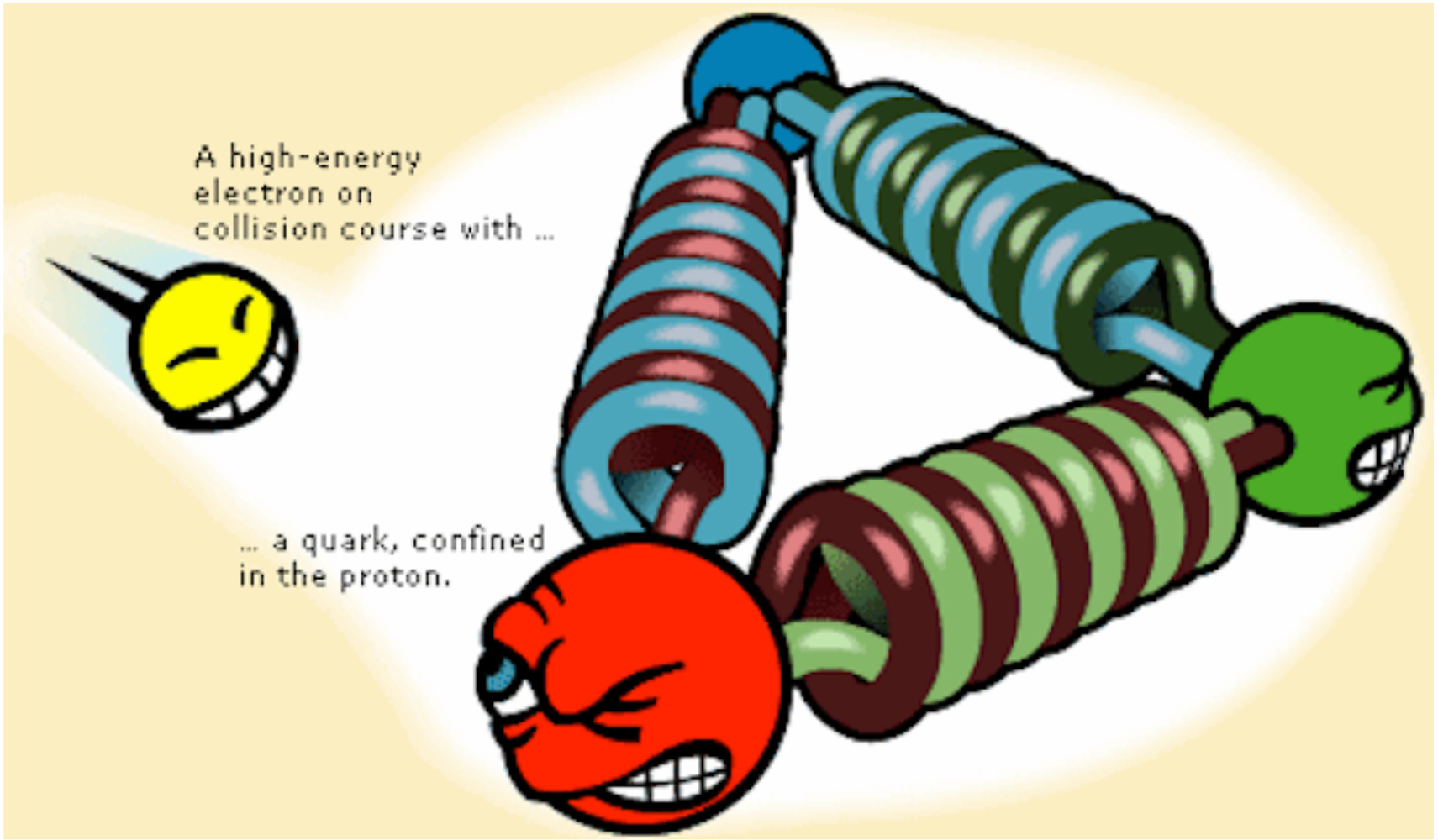
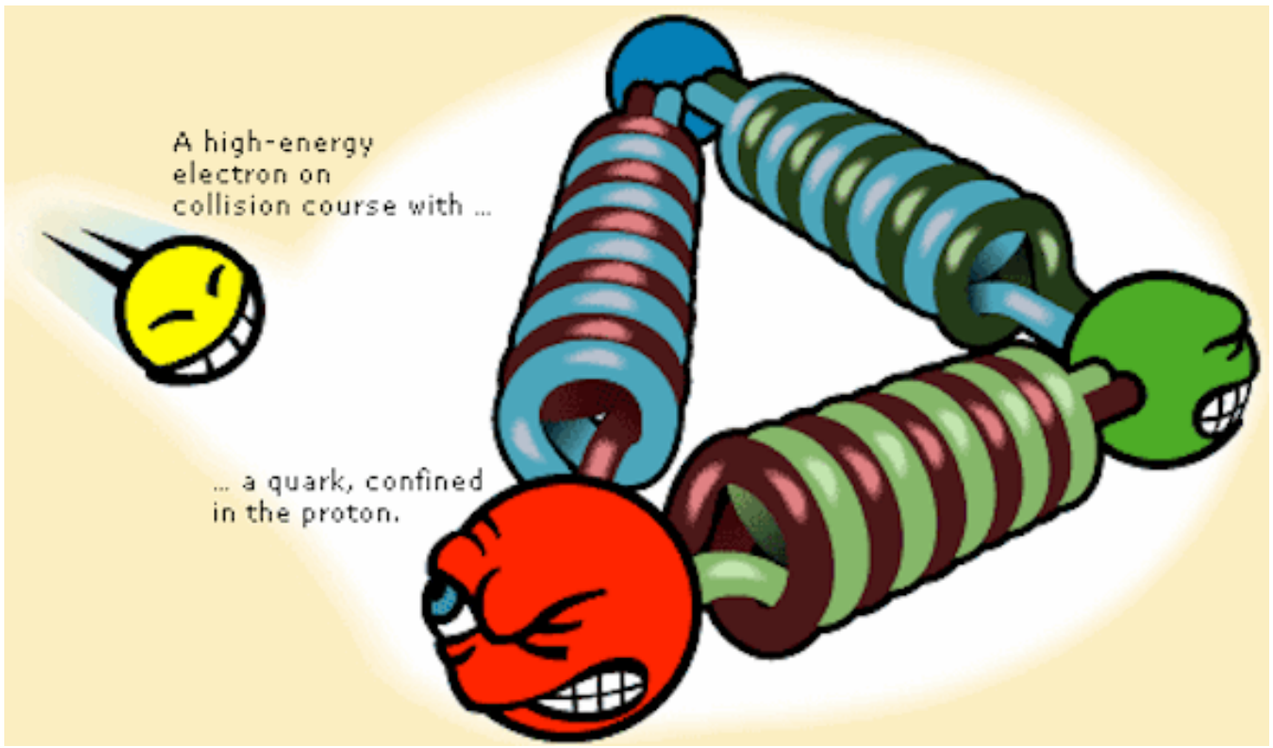


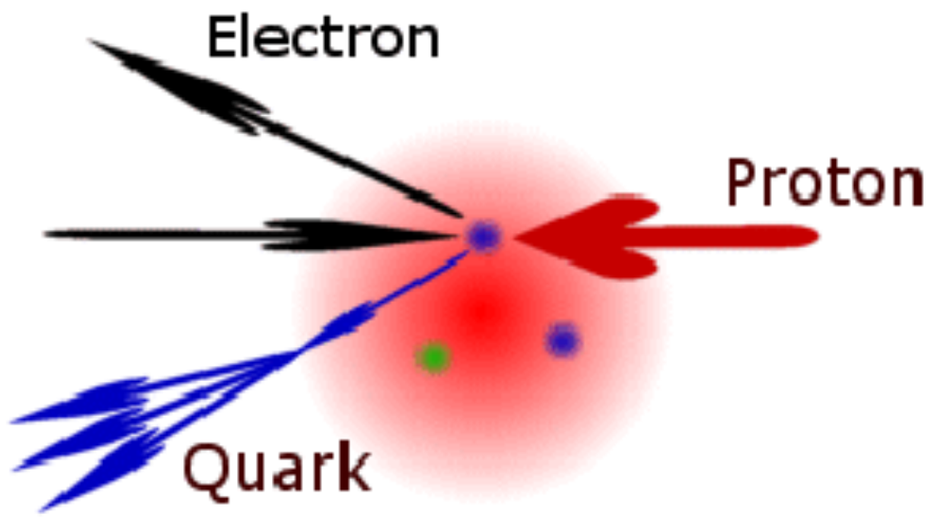
Image taken from the announcement of the winners of the 2004 Nobel Prize in Physics for

"the discovery of asymptotic freedom in the theory of the strong interaction"

- In **Quantum Chromodynamics**, which describes the strong nuclear force, the mathematics says that the closer quarks are to one another, the weaker the force becomes. Conversely the further the quarks move apart, the greater the force is.
- Critical to understanding hadronic matter (e.g. the protons at the **LHC**)
- ***How can this be tested ?***

Theory meets phenomenology

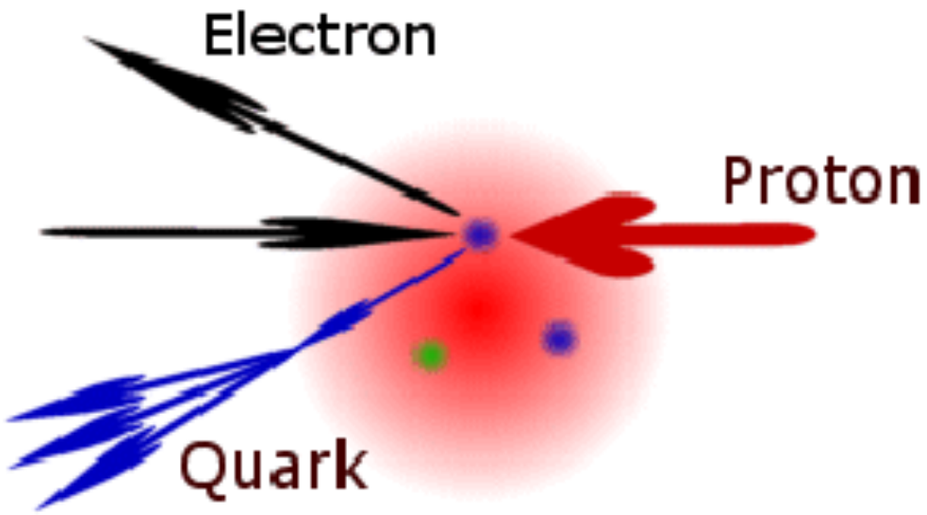


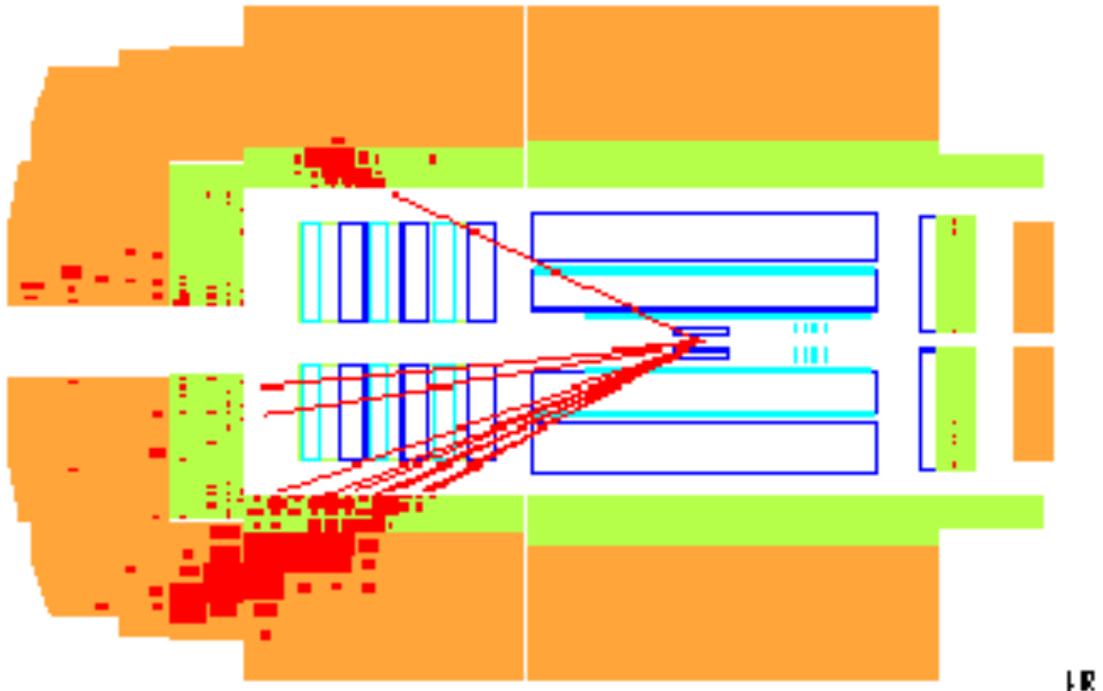


- Test in the old-fashioned particle physicist way
- ***Smash particles together!***

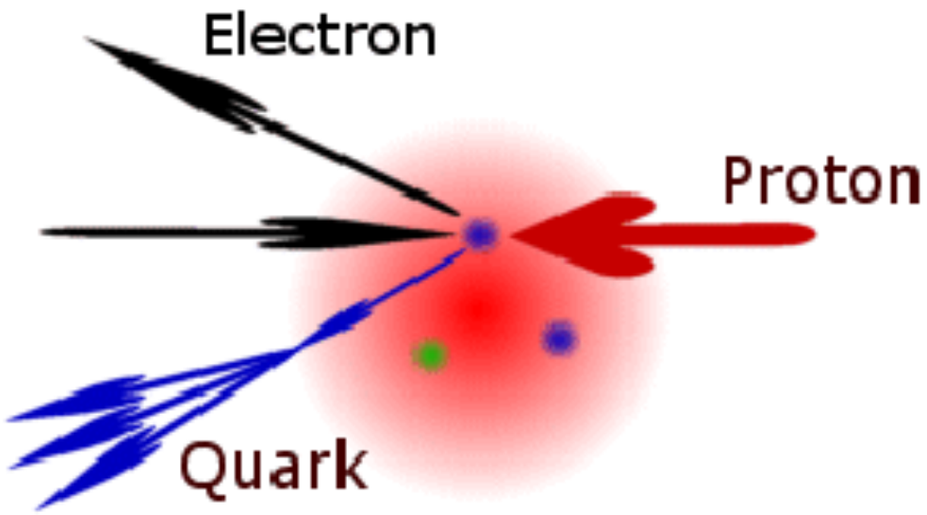
- Model the **proton** as a fuzzy bag of **quarks** and **gluons**, hit it with a high energy (*point-like*) **electron** - what will happen?
- (More on hadronic physics modelling in **Mike Seymour's** lecture)

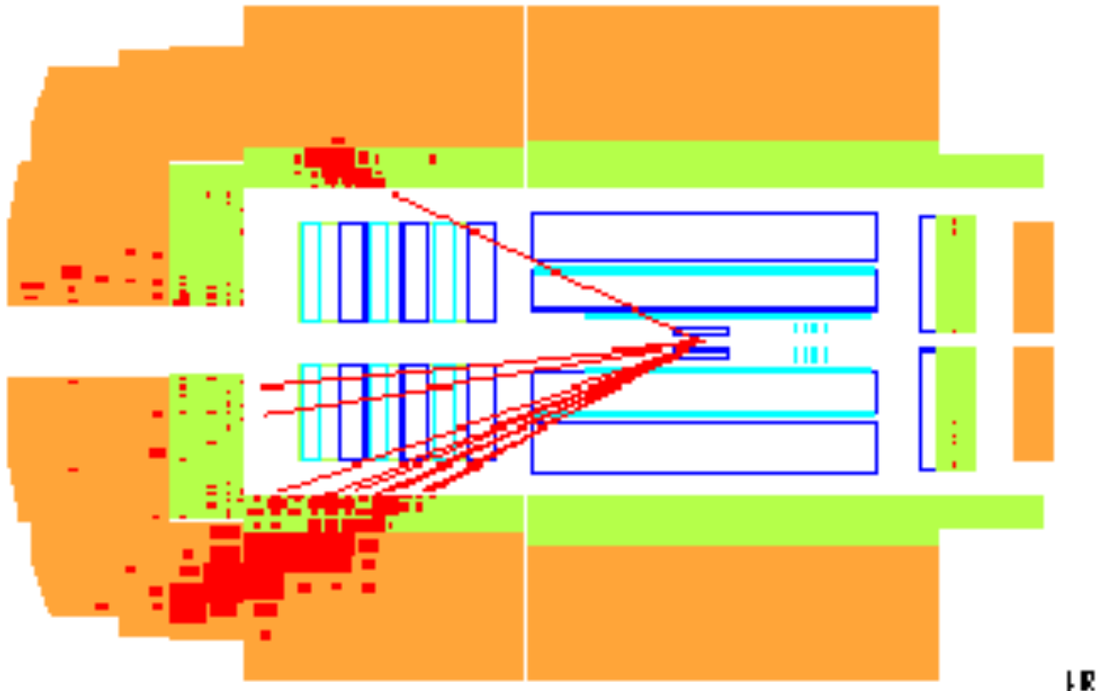
Phenomenology to experiment





Phenomenology to experiment

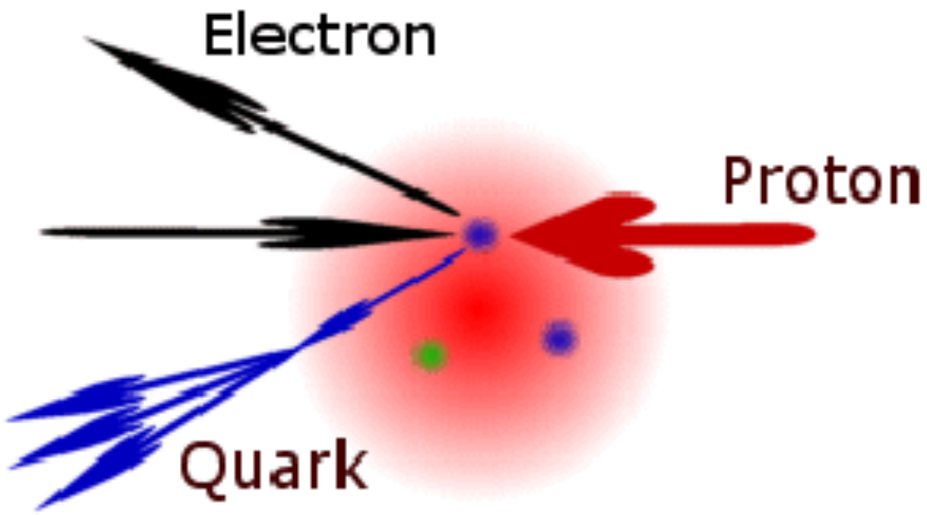


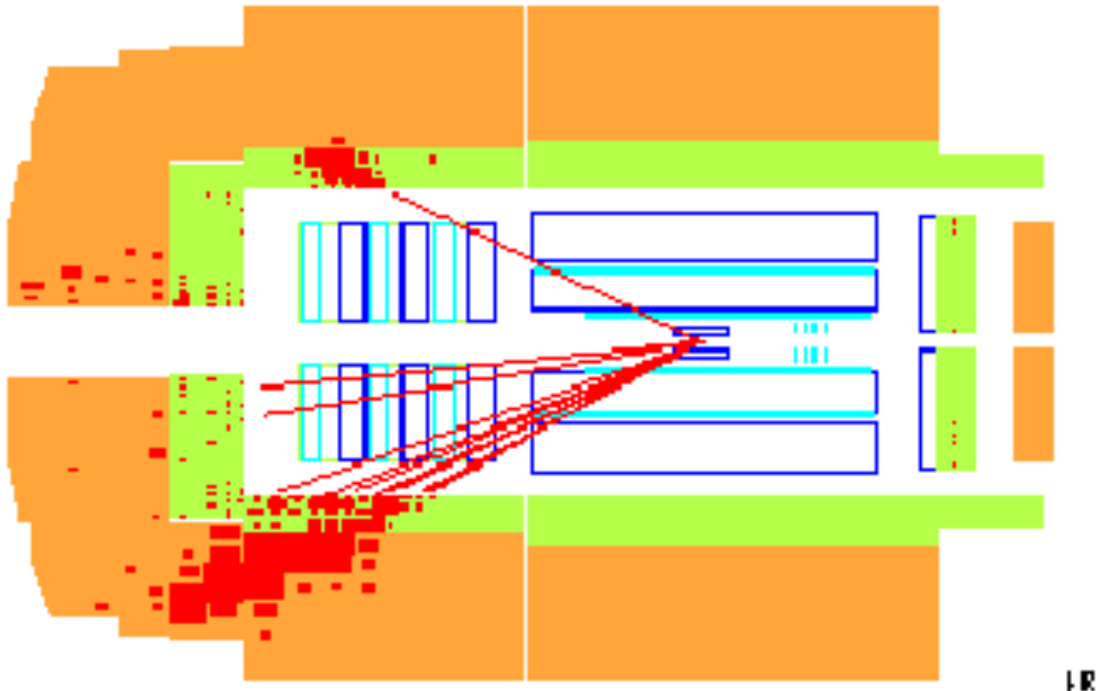


Measure:

$$\frac{\mathrm{d}^2\sigma^{ep}_{NC}}{\mathrm{d}x\mathrm{d}Q^2} = \frac{2\pi\alpha^2 Y_+}{xQ^4}\left(F_2(x,Q^2) - \frac{y^2}{Y_+}F_L(x,Q^2)\right)$$

Phenomenology to experiment





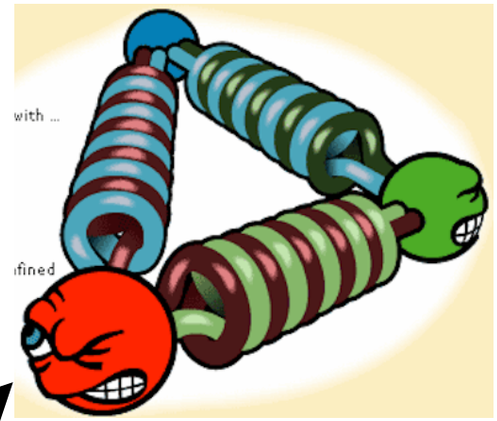
Measure:

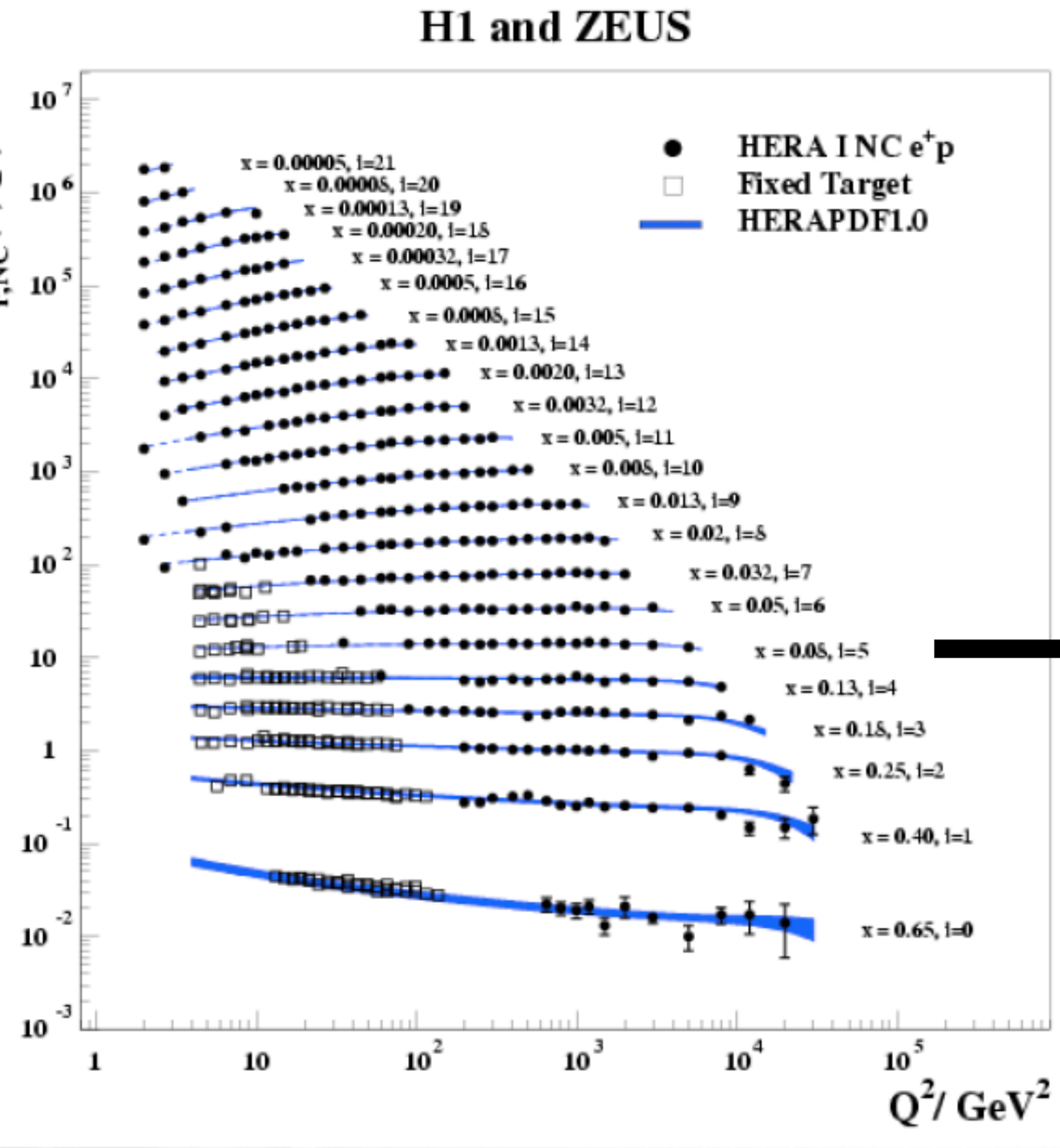
$$\frac{\mathrm{d}^2\sigma^{ep}_{NC}}{\mathrm{d}x\mathrm{d}Q^2} = \frac{2\pi\alpha^2 Y_+}{xQ^4}\left(F_2(x,Q^2) - \frac{y^2}{Y_+}F_L(x,Q^2)\right)$$

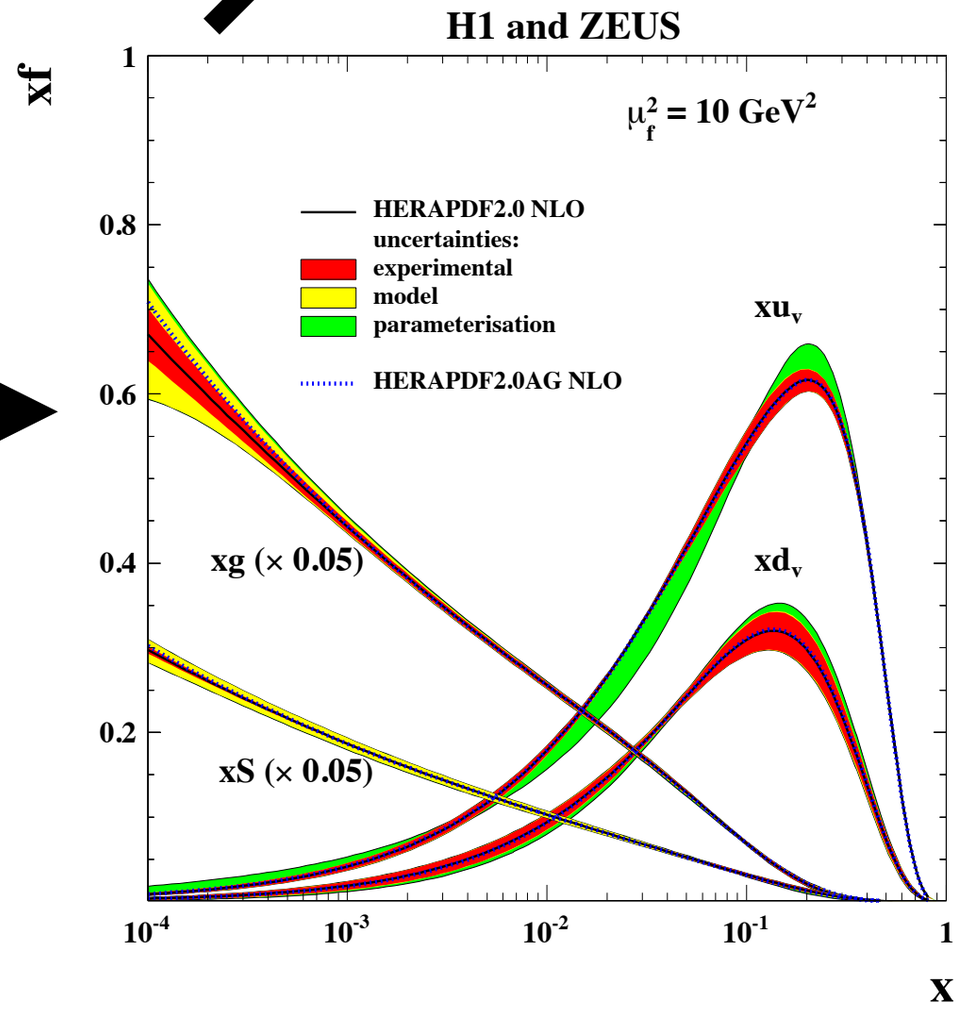
Extract:

- F_2 directly related to (PDFs) quark content: $F_2 \sim x\Sigma e^2(q+q)$
- $dF_2/d\ln Q^2$ (scaling violations) sensitive to gluon content
- F_L only non-zero in higher order QCD – independent access to gluon density and QCD dynamics

Extracting observables





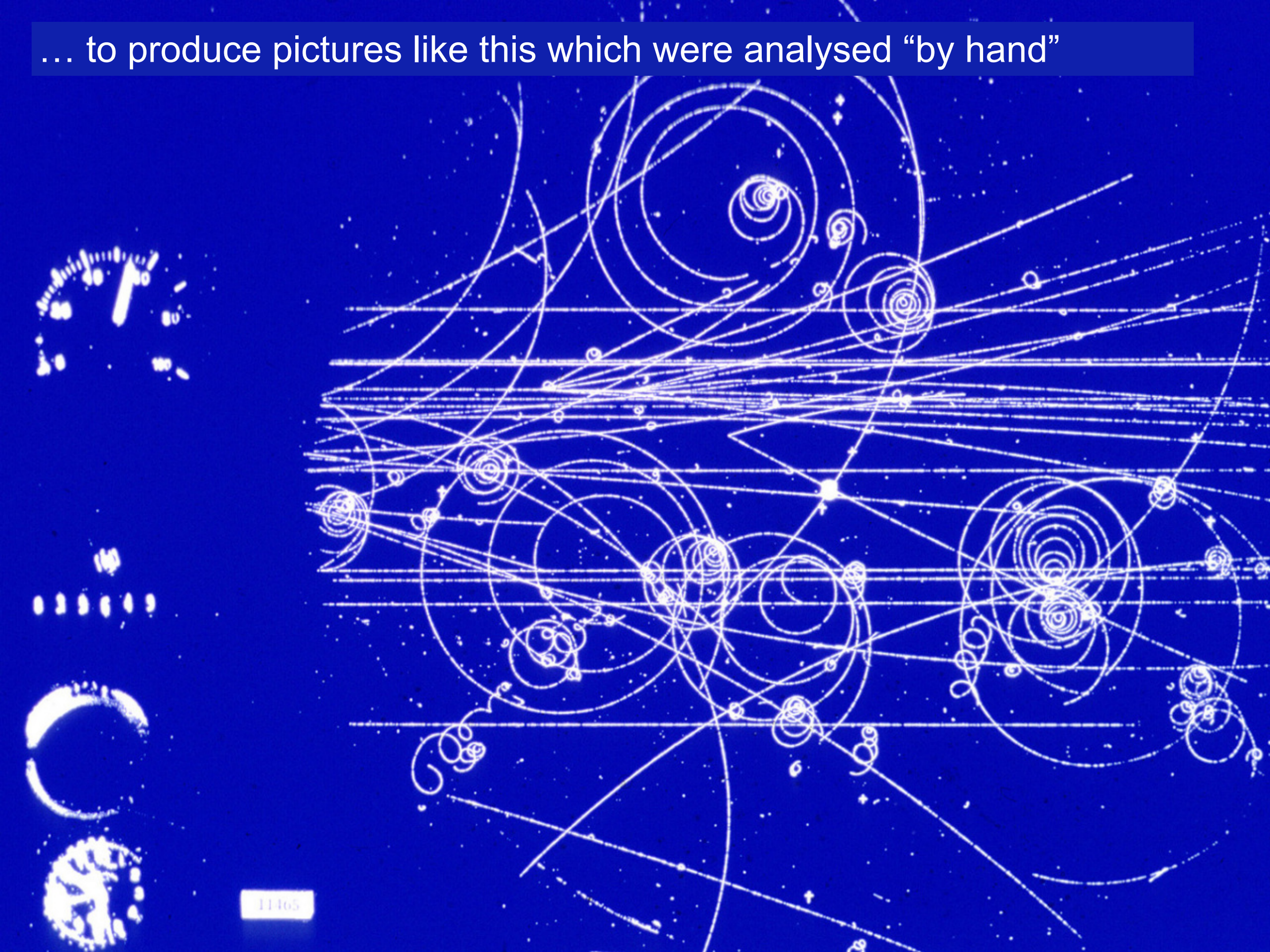


- Experimental confirmation of **QCD** allowing us to infer the **quark** and **gluon** structure of the **proton**

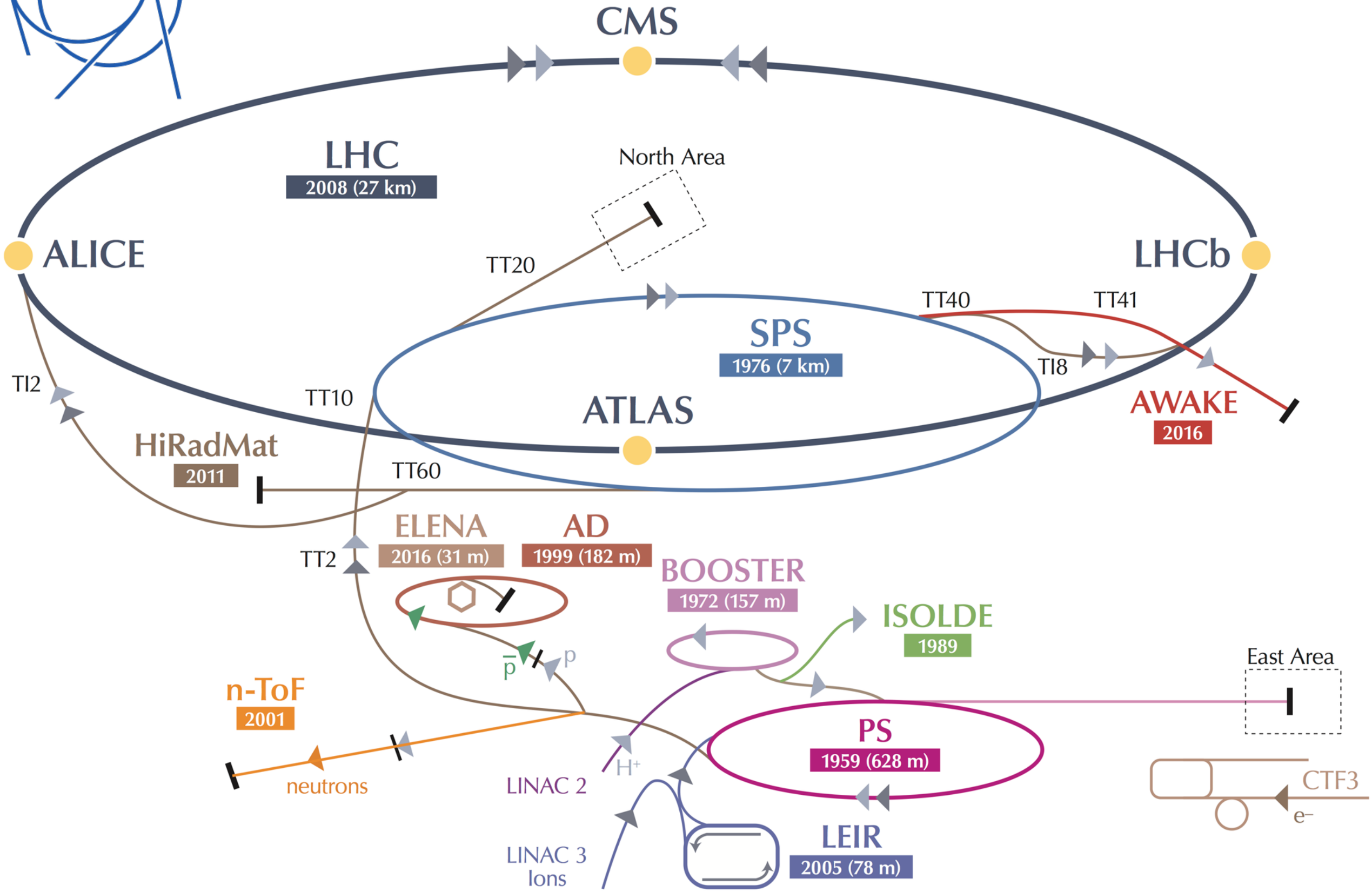
Experiments at CERN

- In the 1960s we used Bubble chambers, the one that you can see in the Microcosm was used…

… to produce pictures like this which were analysed “by hand”



Today @ CERN we have huge rates of collisions so that we can produce very rare events



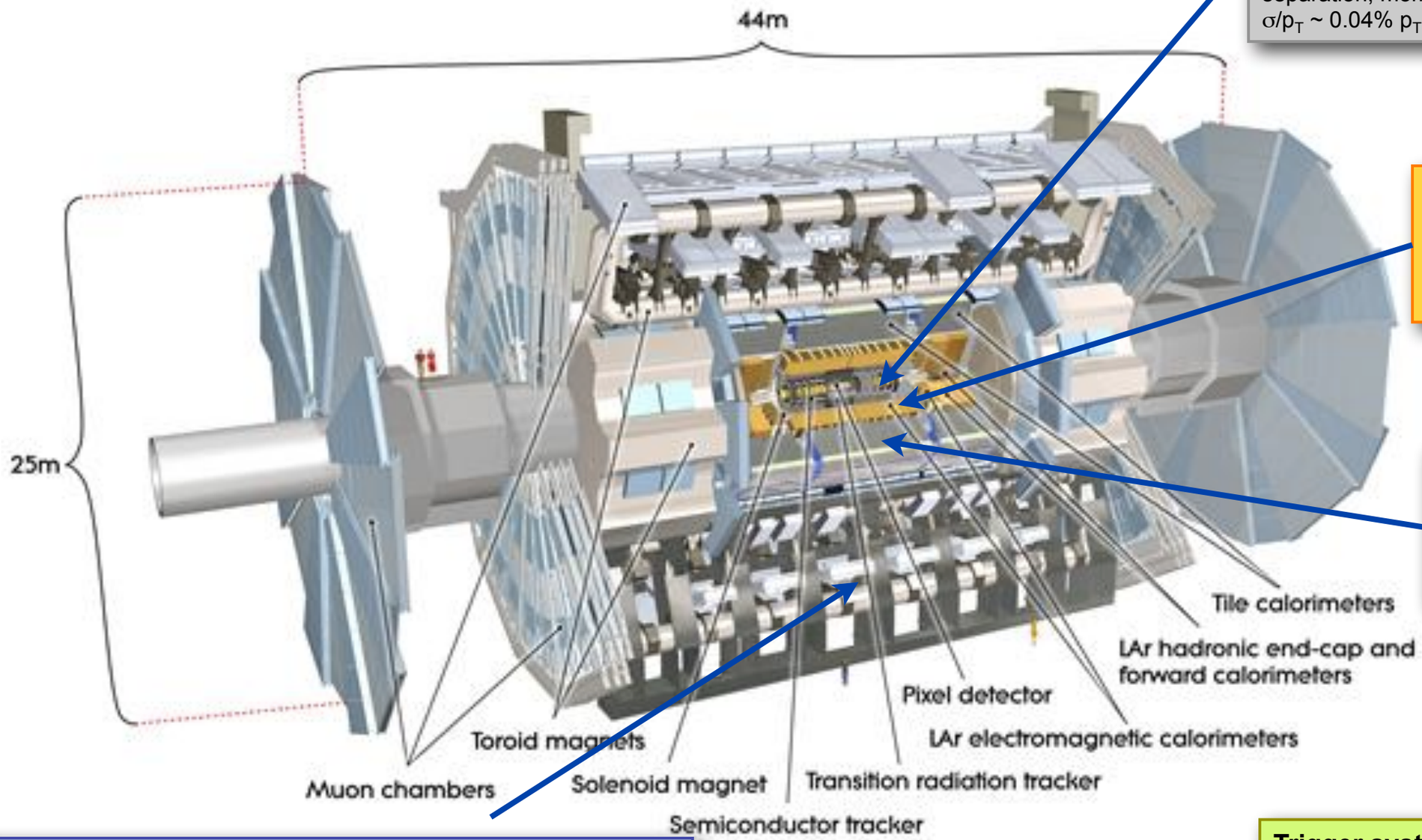
The ATLAS Detector @ LHC

L ~ 46 m, ∅ ~ 22 m, 7000 tons
~10^8 electronic channels

Inner Tracker (|η|<2.5, B=2T): Si Pixels, Si strips, Trans. Rad. Det. Precise tracking and vertexing, e/π separation, momentum resolution: σ/p_T ~ 0.04% p_T (GeV) ⊕ 1.5%

EM calorimeter: Pb-LAr Accordion, e/γ trigger, id. and meas., energy res.: σ/E ~ 10%/√E ⊕ 0.7%

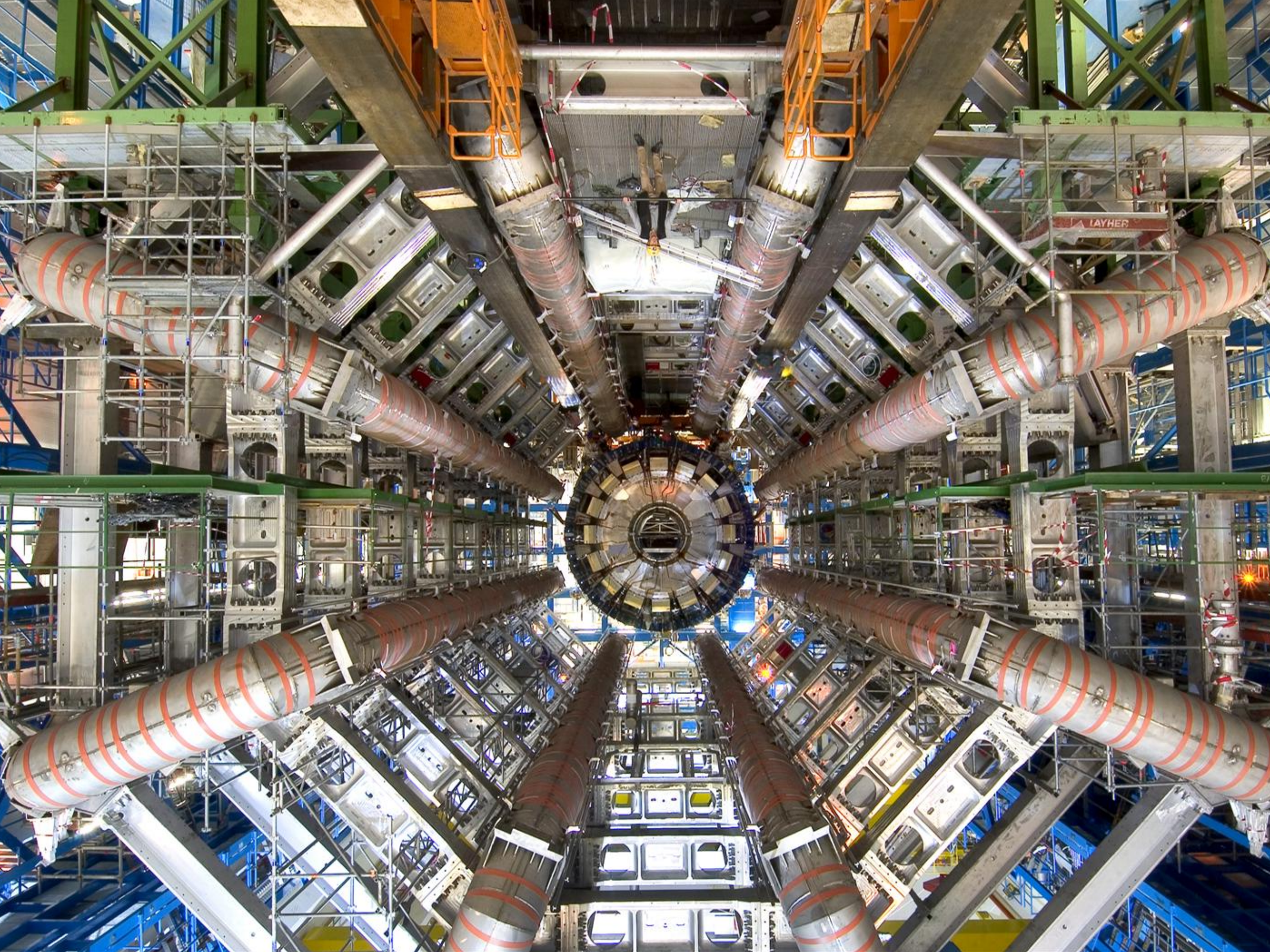
HAD calorimetry (|η|<5): Fe/scintillator Tiles (cen), Cu/W-LAr (fwd). trigger and meas. of jets and $E_{T,miss}$, energy res.: σ/E ~ 50%/√E ⊕ 3%

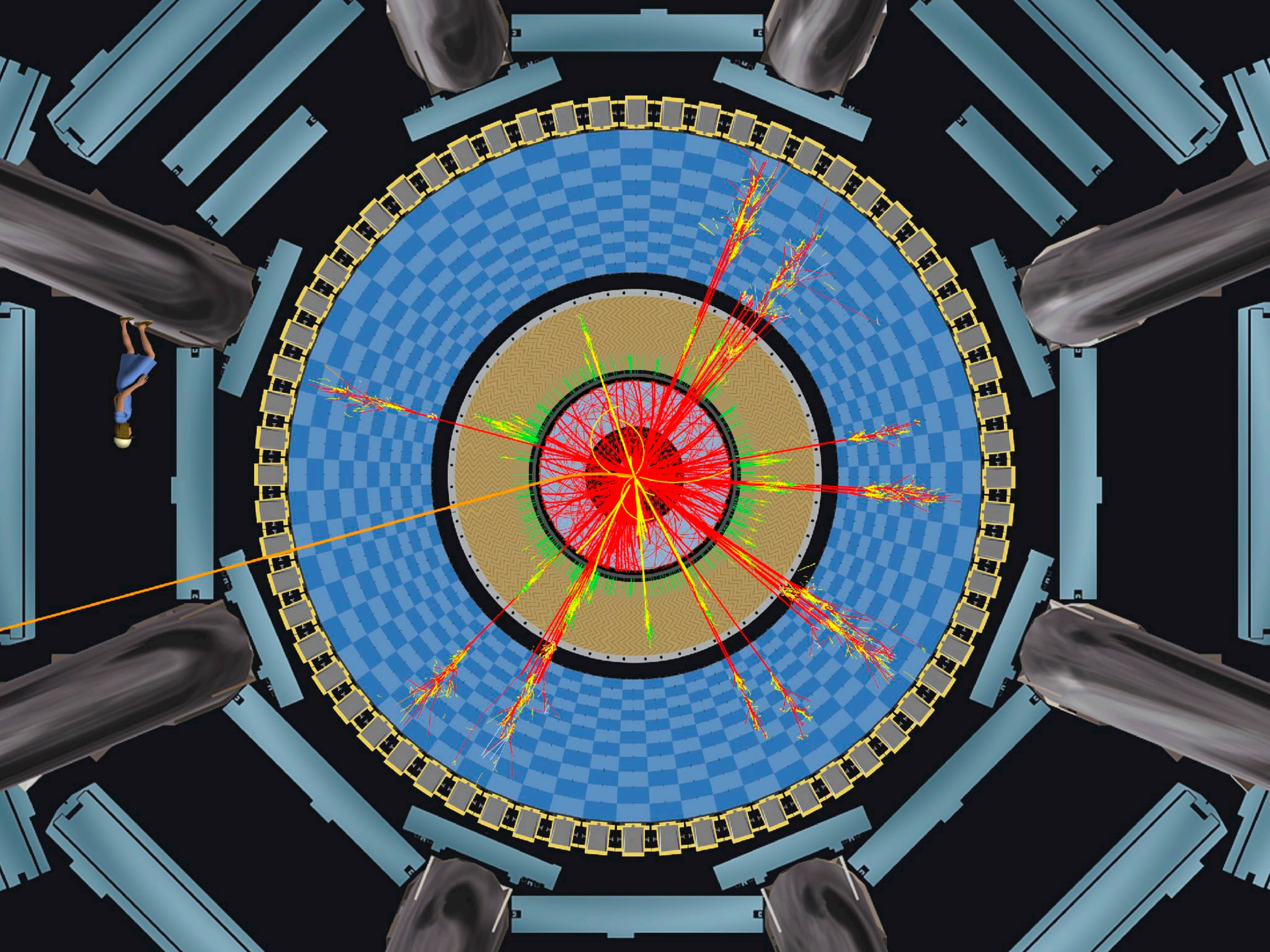


Muon Spectrometer: air-core toroids with gas-based muon chambers. trigger and meas. with momentum resolution < 10% up to E_μ ~ 1 TeV

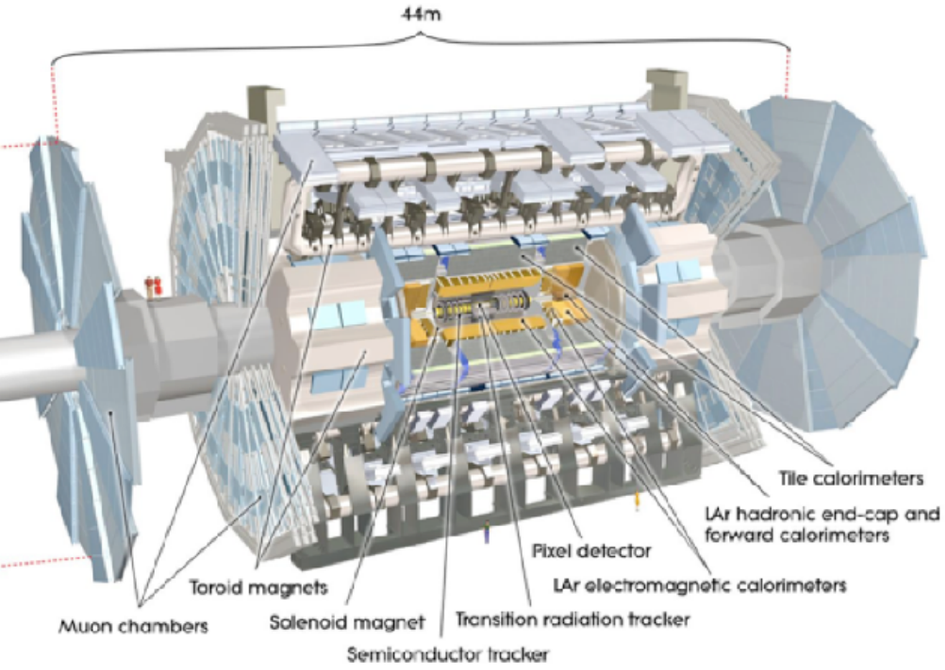
Trigger system: 3-levels reducing the IA rate from 40 MHz to ~200 Hz

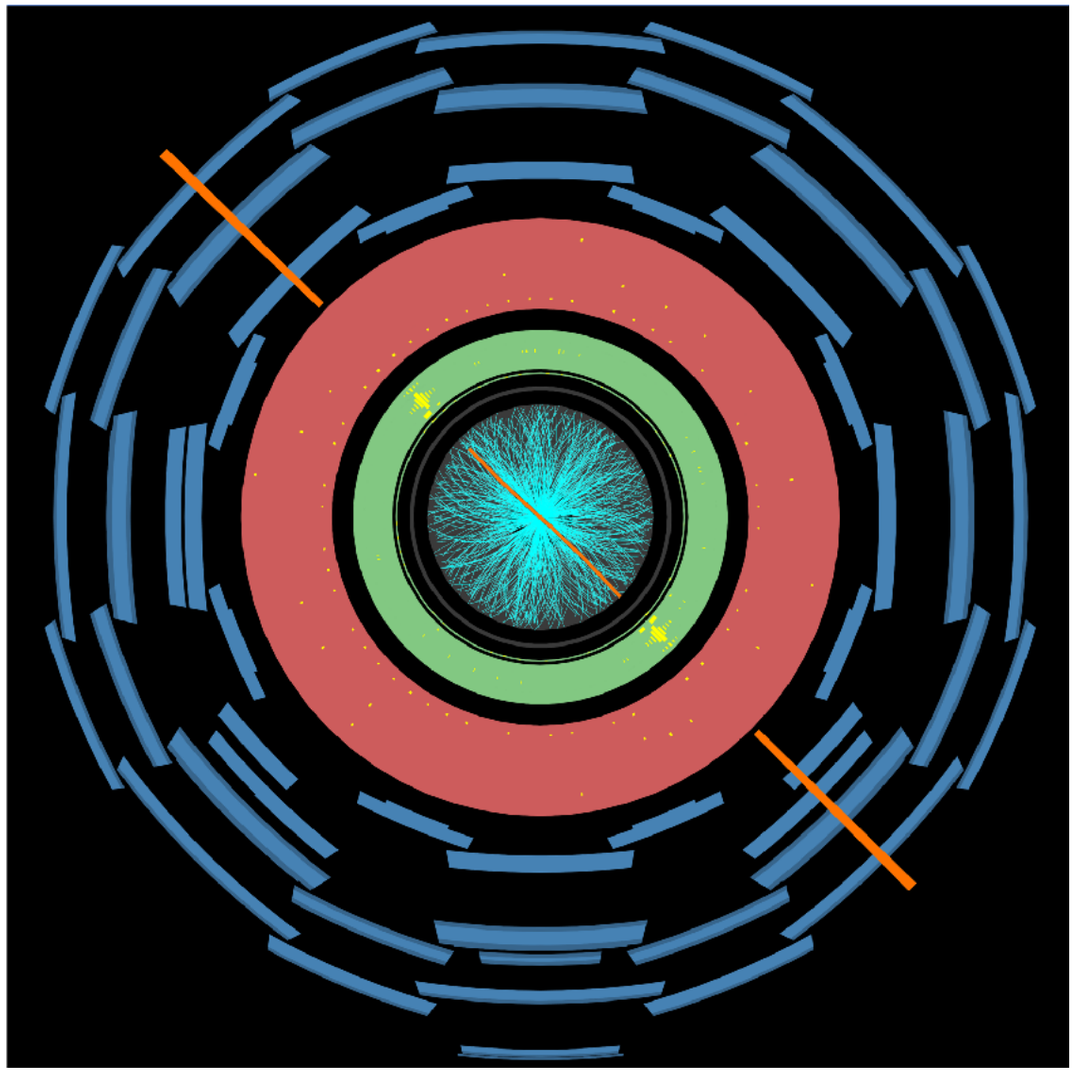
Millions of detector readout channels read out to reconstruct one "event"


LAYHER



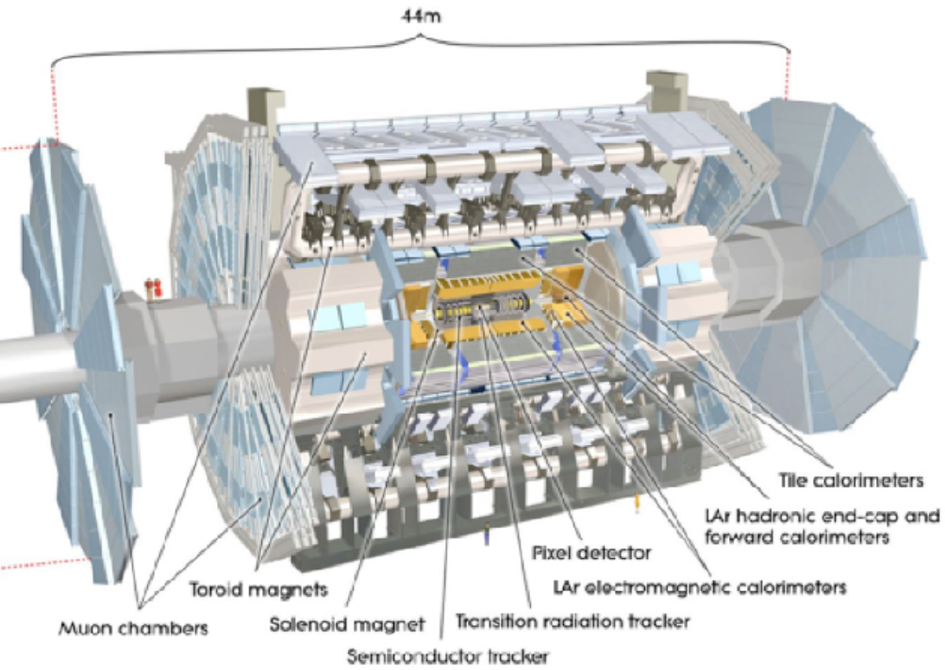
Event displays

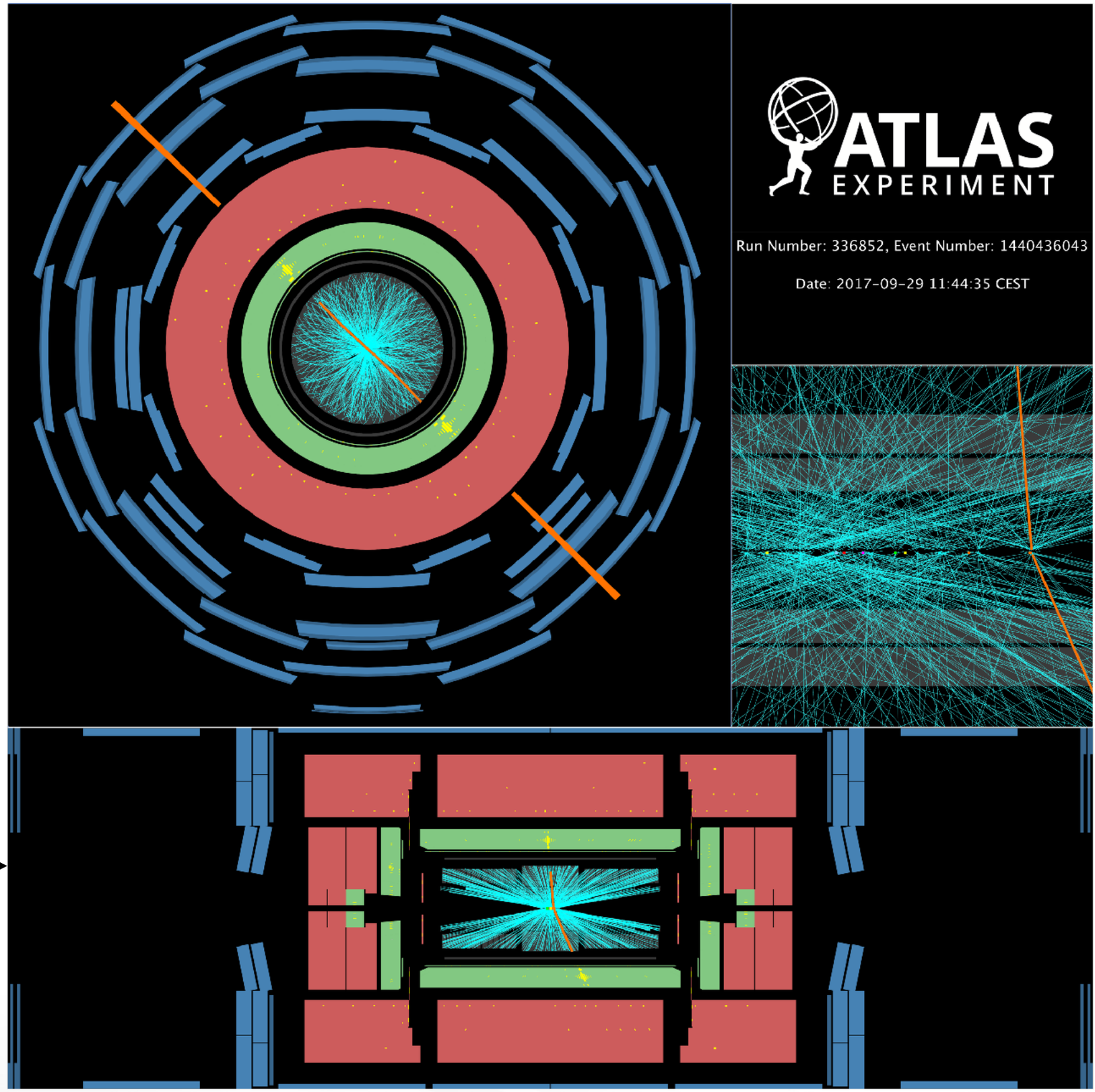




- Event displays are great ways for us to visualise what happened in a particle collision
- In this **ATLAS event display** (*right*) of a real proton collision, we are looking down the beam pipe, so the plane of the display is transverse to the proton beam direction
- ***Question****: Can you quantify the momentum in this plane* ***before*** *the proton collision*
  - *What does that tell you about the distribution of momentum* ***after*** *the collision?*
  - *Can you say which fundamental particle(s) is (are) observed in the event?*

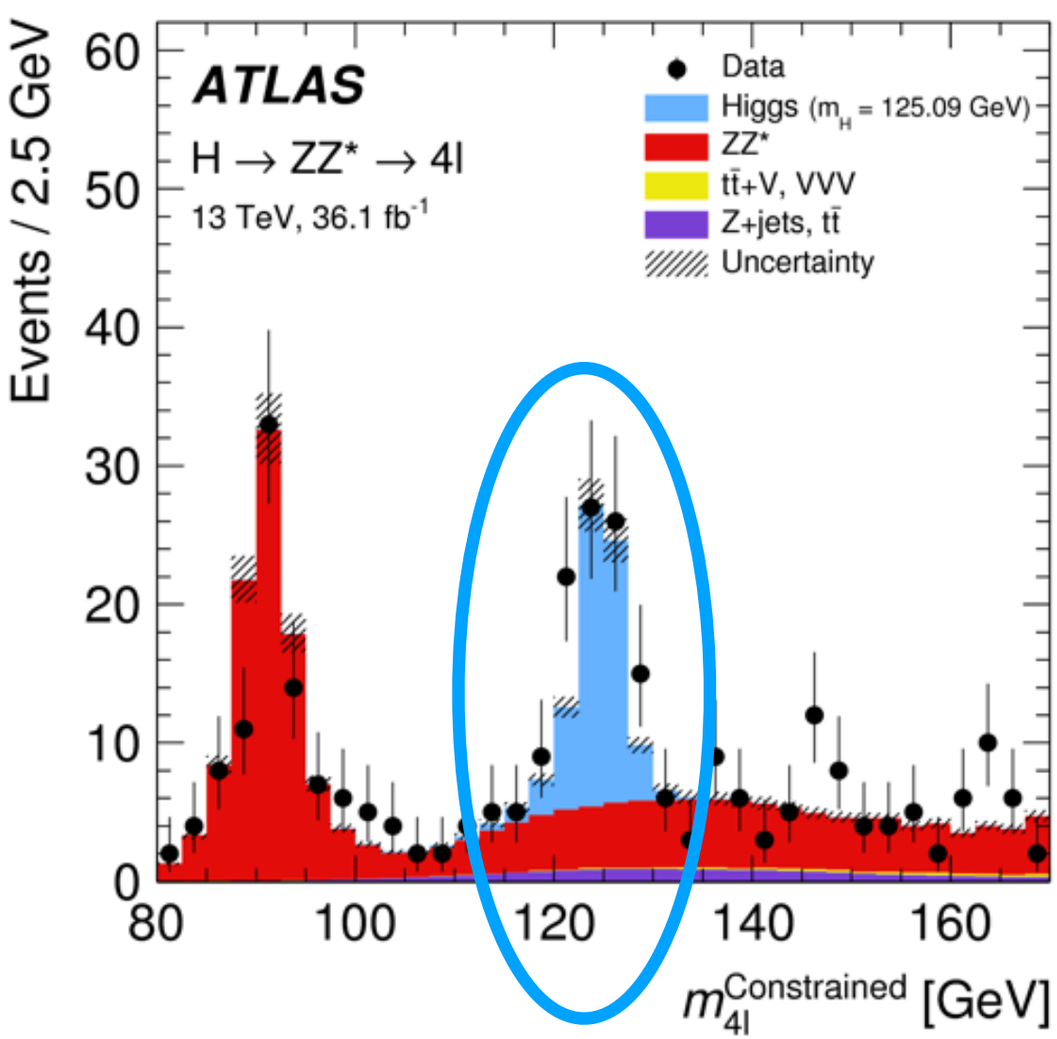
Event displays

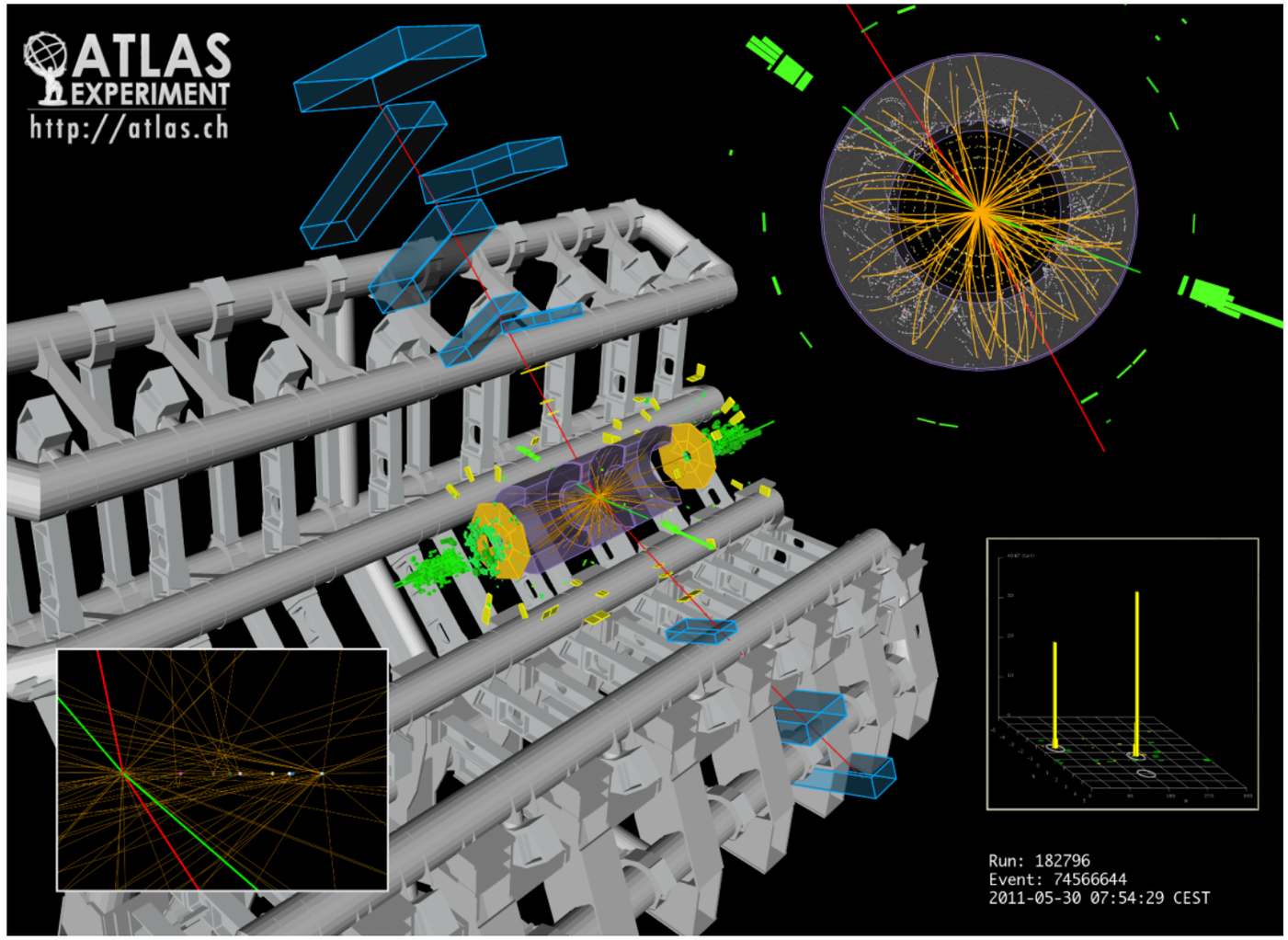




- This view shows the plane in the proton beam direction →
- Both **2D** views are often used to provide complementary information

Discovering the Higgs Boson

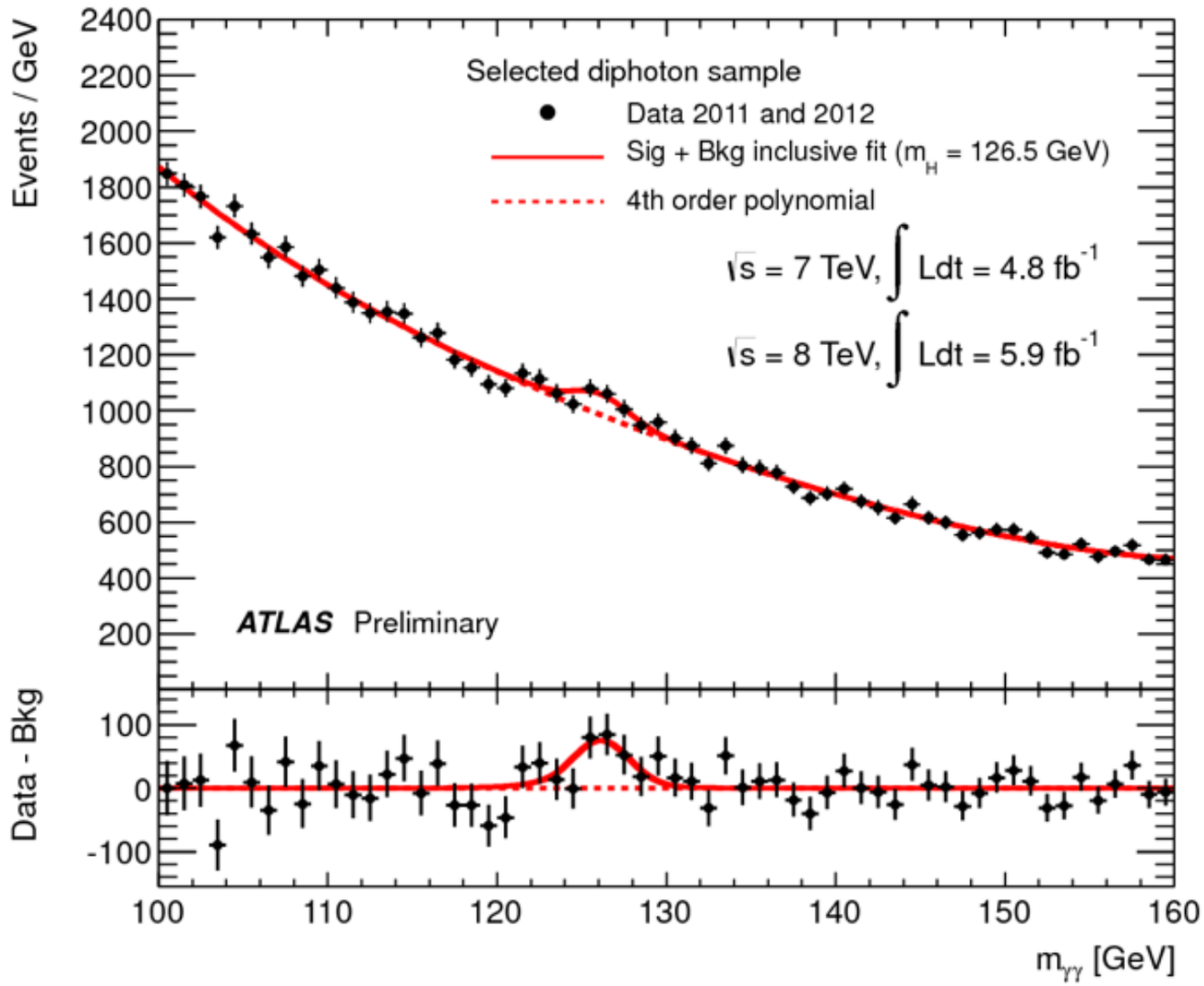




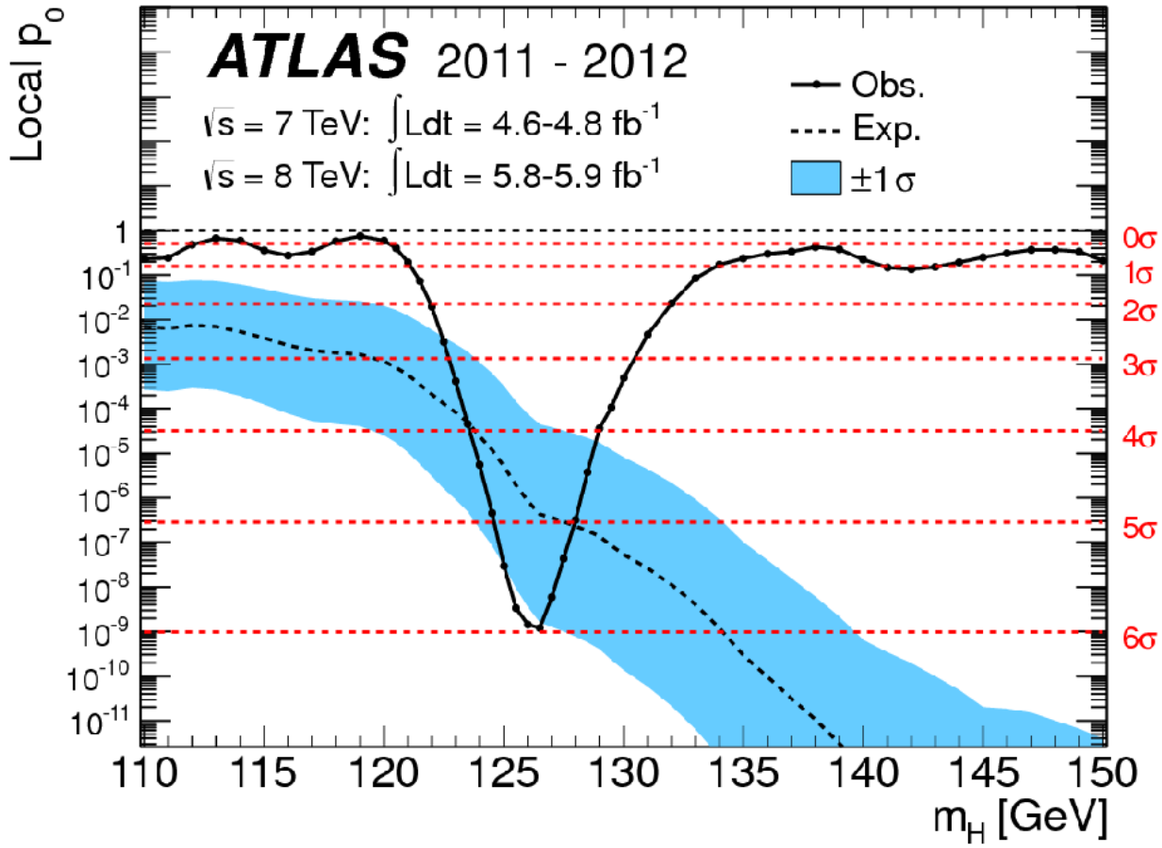
- Strategy for this channel (there are several) - we look for events with **two Z bosons** that have decayed to **four leptons**, e.g. two electrons and two muons in the event on the right
- If the **two Z bosons** were produced by the **decay of a Higgs boson**, when we reconstruct the invariant mass of the system we should see a **peak** at the **Higgs boson mass**

Needles in haystacks

- There are billions of events and the ones we are really interested in are ***very rare***
- Often the interesting events are also ***very difficult to distinguish*** from background
  - Requires ***high precision detectors***, which means ***lots of data*** for each event
- The data are structured but each event is different - ***unique data science challenge***



Higgs discovery in 2012



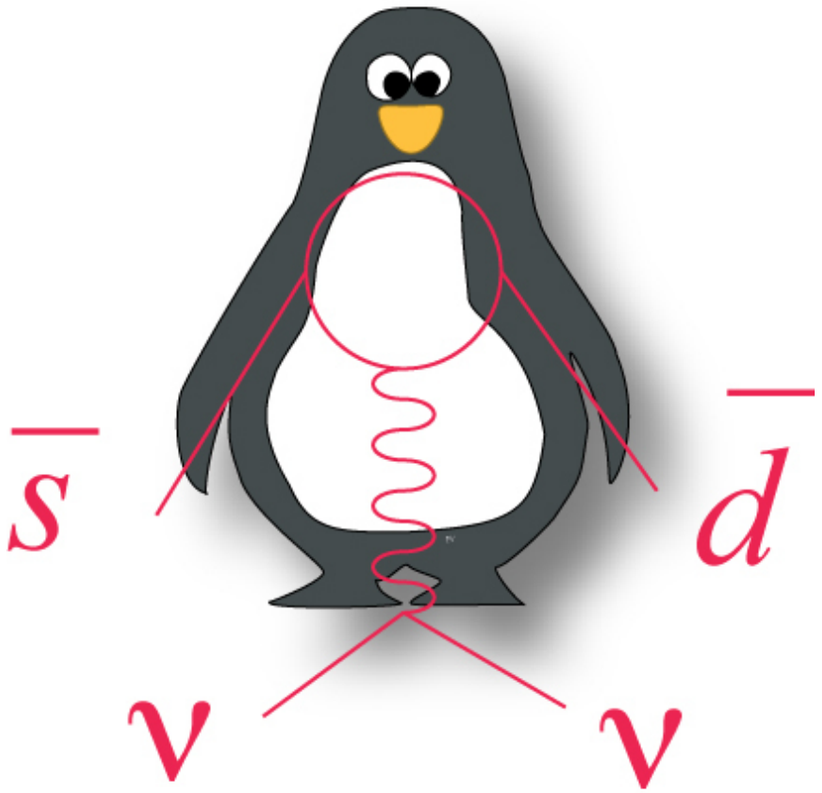
- In 2012 the number of observed events (**6σ**) was consistent with, and in excess of the number of events expected for a standard model Higgs (**5σ**)

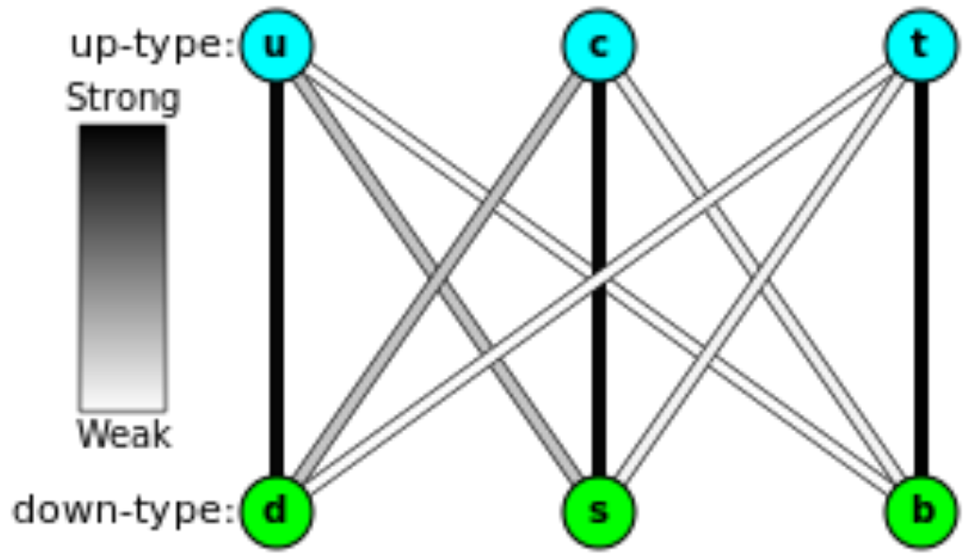
- ***Question*** *- Imagine we had several more Large Hadron Colliders, with a total of 9 independent measurements possible. Roughly how many measurements would you expect to lie* ***outside*** *the ±1σ blue band?*

Penguin Experiments

NA62

How strange quarks turn into other quarks





$$BR(K^+ \to \pi^+ \upsilon\bar{\upsilon}) = (8.4 \pm 1.0) \times 10^{-11}$$ [Buras et al. JHEP 1511 (2015) 33]

Theory predicts this happens ~100 times in a million million kaon decays:

1) Record a million million kaon decays
2) Analysis - do we count 100 signal events?
   Yes - Congratulations, that's very impressive!
   No - *Congratulations, you've discovered new physics!*

1 in 10 billion is the same probability as finding you in the Earth's population

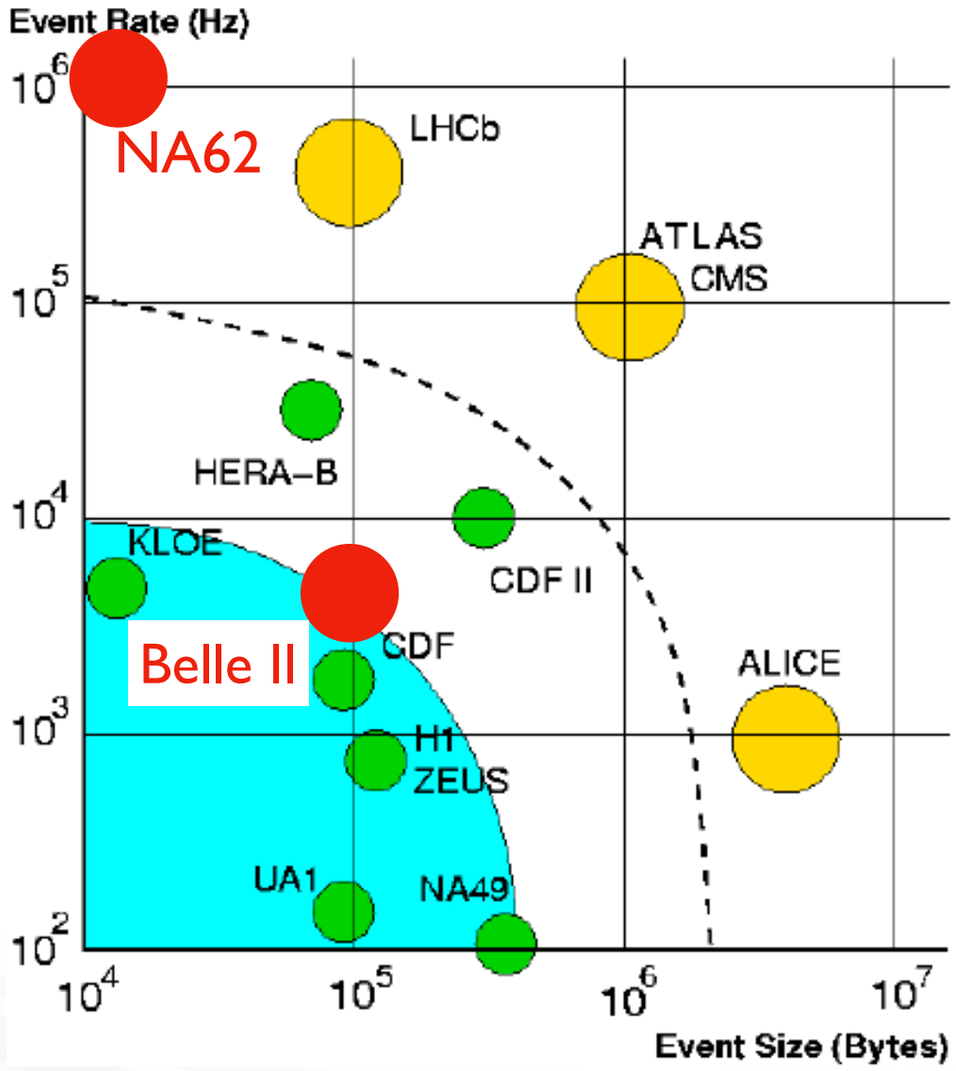
Raw data throughput

Plot modified from:
"GridPP: development of the UK computing Grid for particle physics."

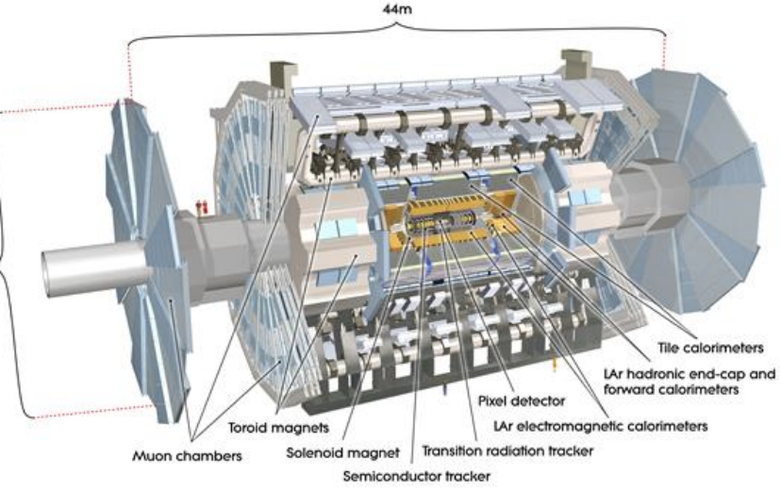
- H1
  - Proton structure and QCD
  - small event sizes and rates
- ATLAS
  - Higgs, searches for new physics
  - big event sizes and rates
- NA62
  - Ultra-rare kaon decays
  - huge rates of small size events
- Belle II
  - Ultra-rare B decays
  - modest event sizes and rates

- Triggers are critical to reduce the amount of data we record and analyse later

DAQ throughput =
event rate * event size



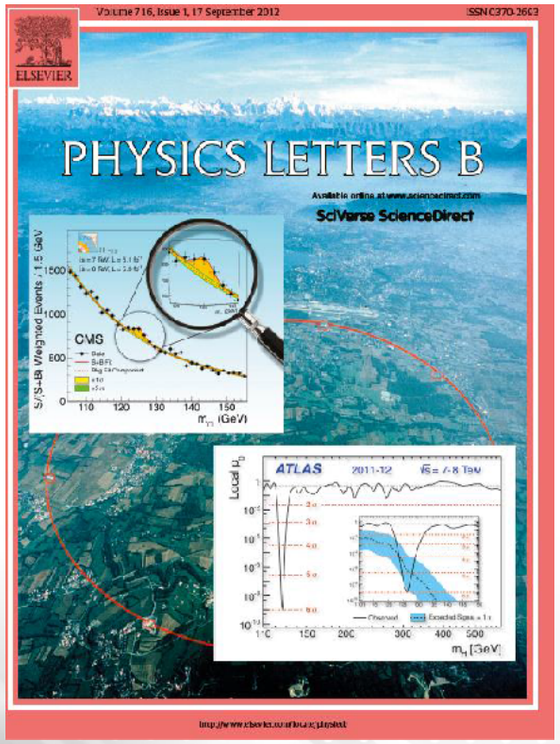
Data's journey



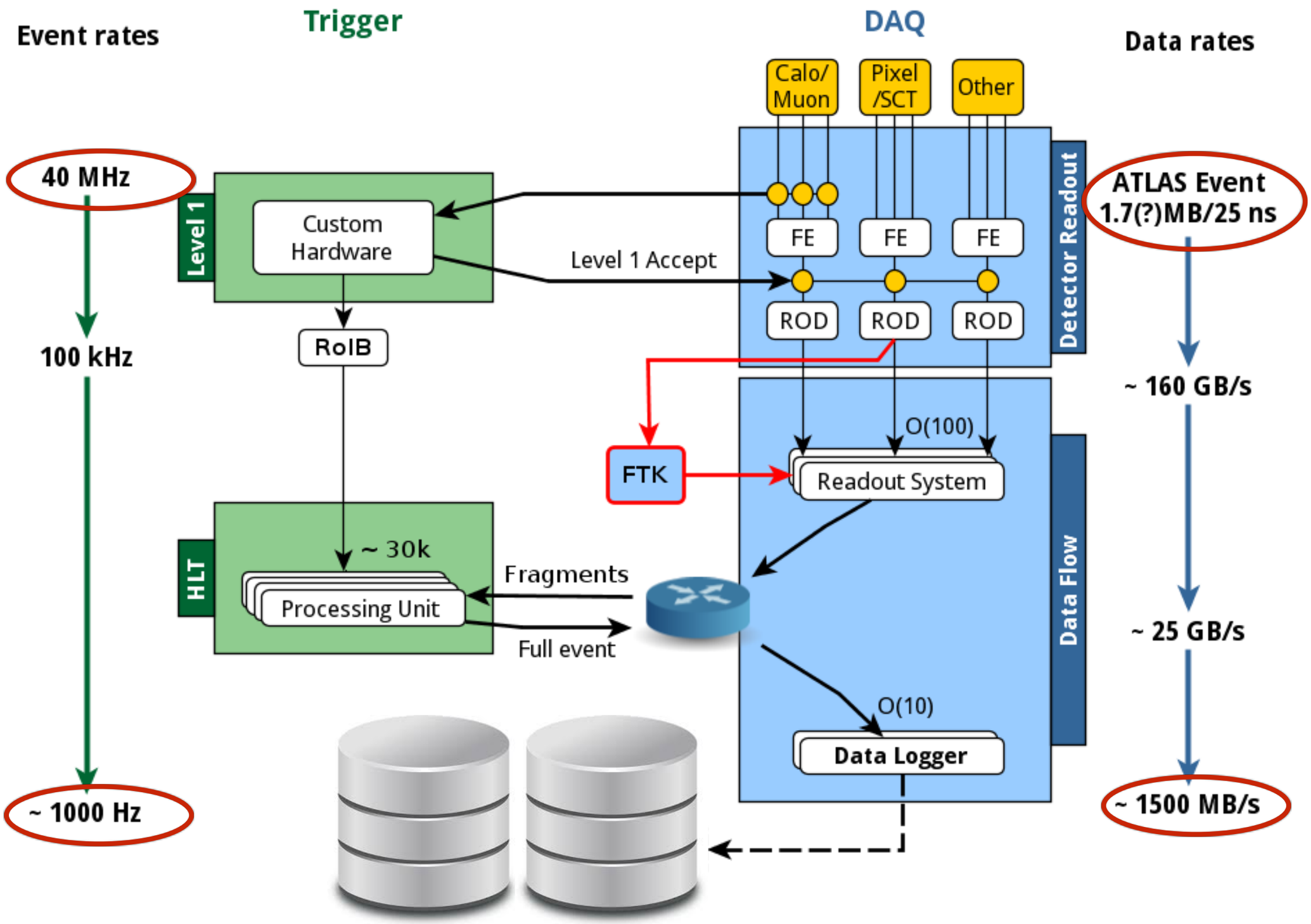
Trigger/DAQ

Data Preparation

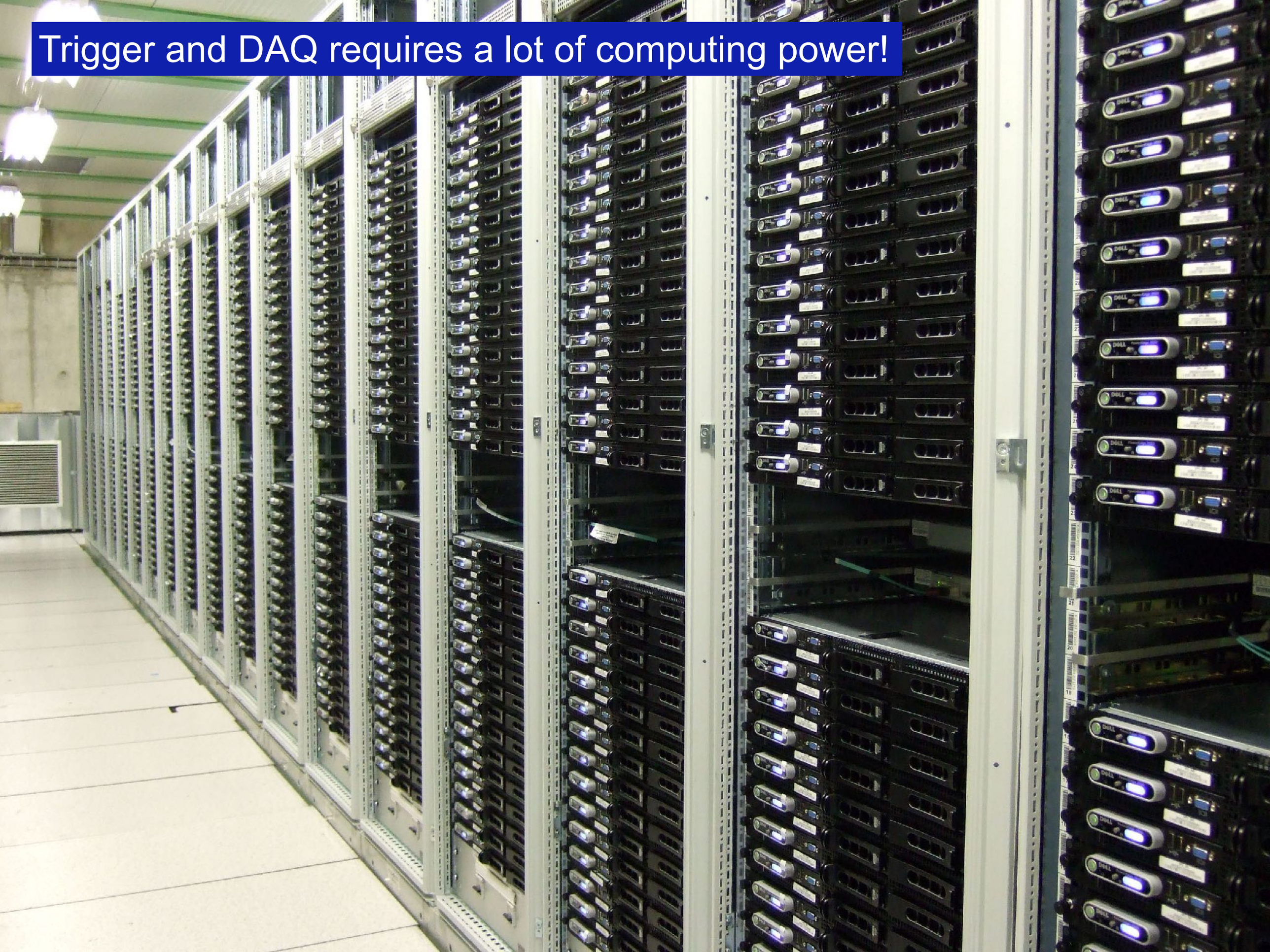
Distributed computing



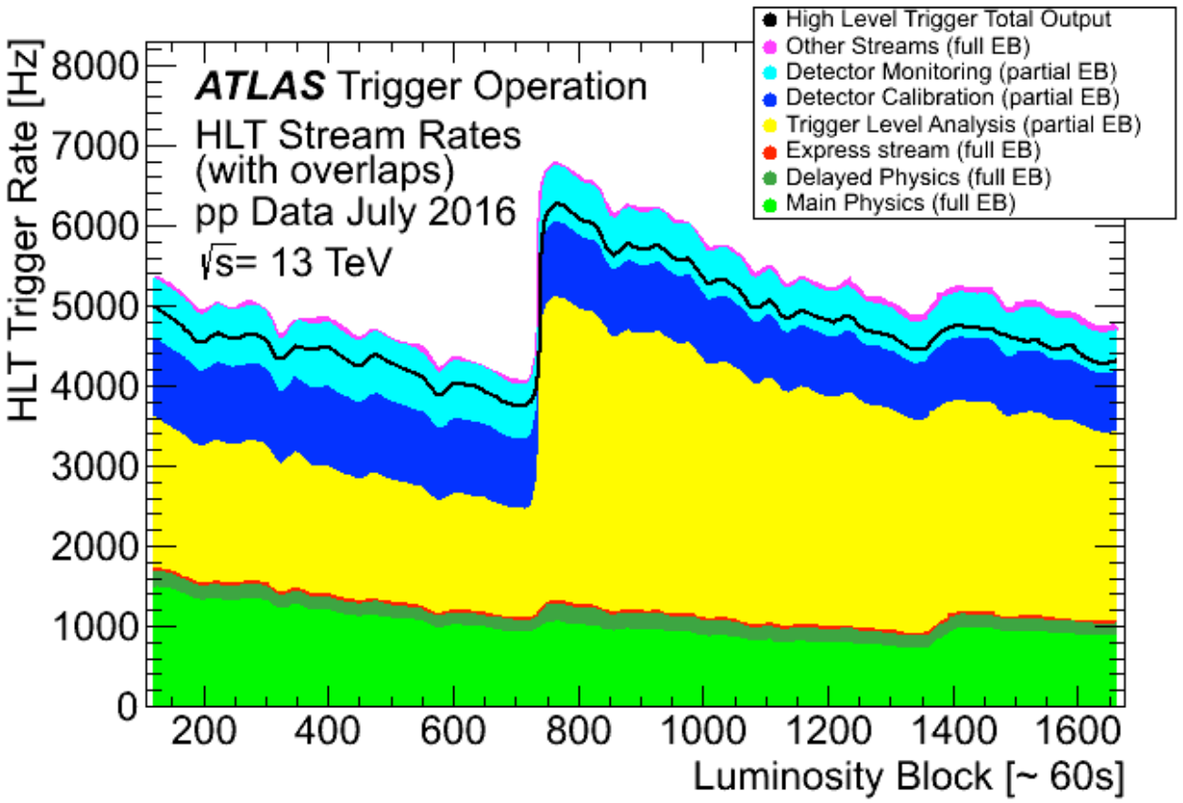
The Atlas Trigger and DAQ

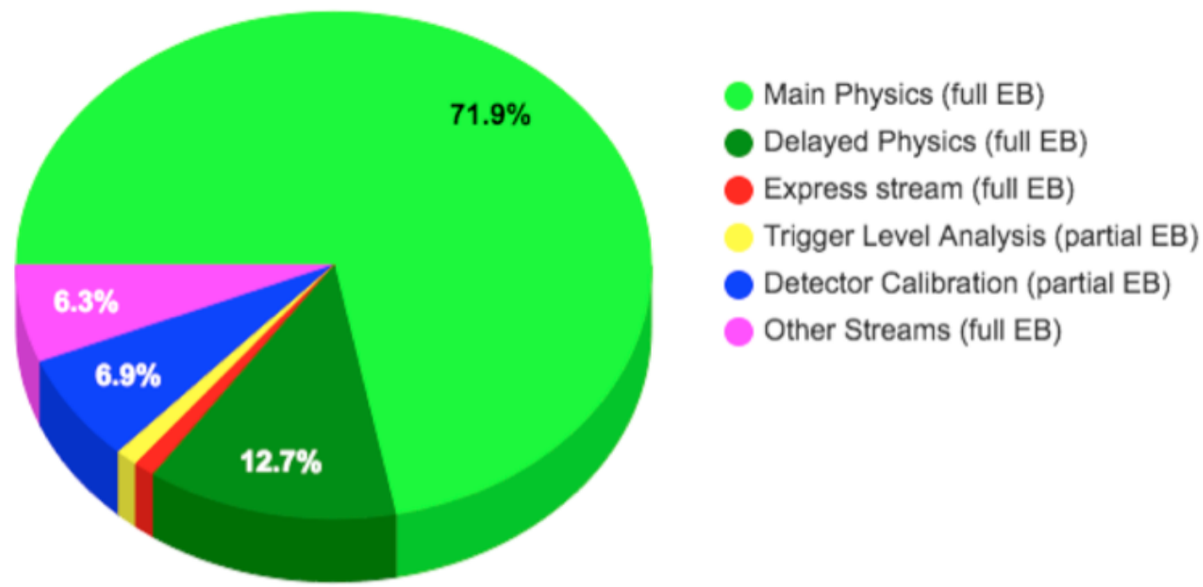


Trigger and DAQ requires a lot of computing power!



Trigger streams





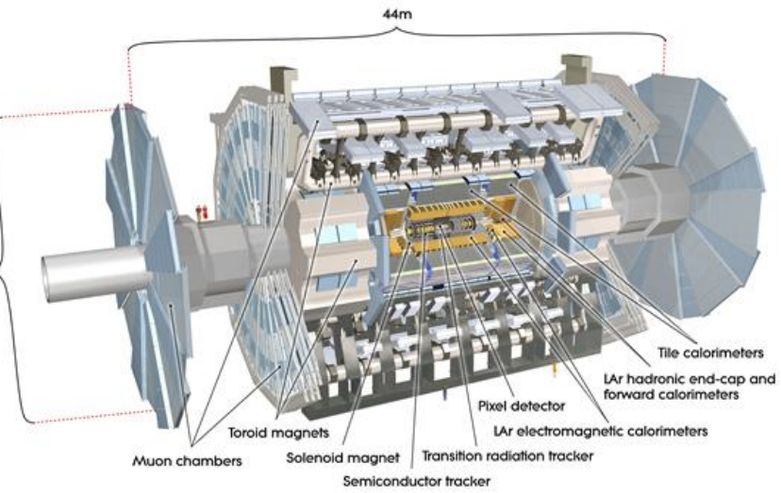
- We know in advance that in addition to the main physics data, we will need some **dedicated data** for:
  - *performing calibrations*
  - *assessing data quality*
- Writing **dedicated output streams** (written to different physical files) provides people with just the data they need

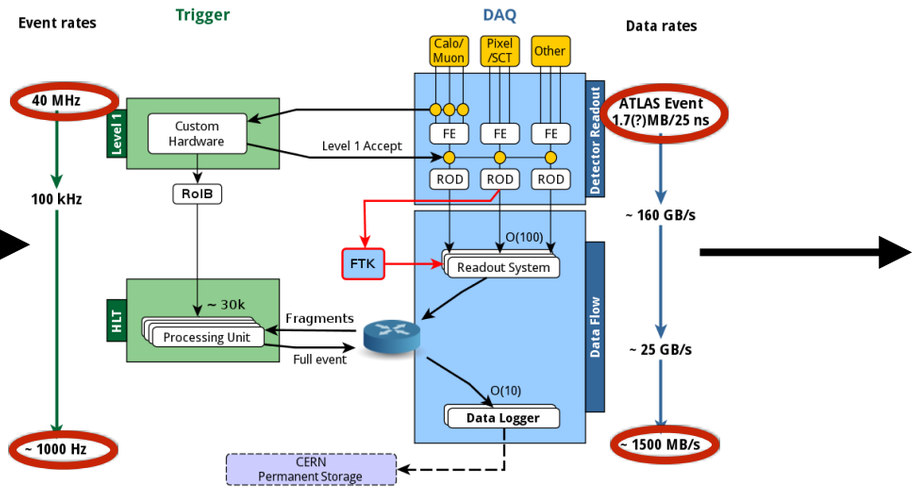
What happens to the data?

- You have learned about the *detection of particles*
  - How *different particles interact with different detectors* - **Detectors**
  - How this is a *statistical process* - **Foundation of Statistics**
  - How the results of the particle interactions leave deposits and signals in the detectors that we *read out* - ***Electronics, DAQ and Triggers***
    - We read out detectors when a ***trigger*** *flags the event as being interesting*

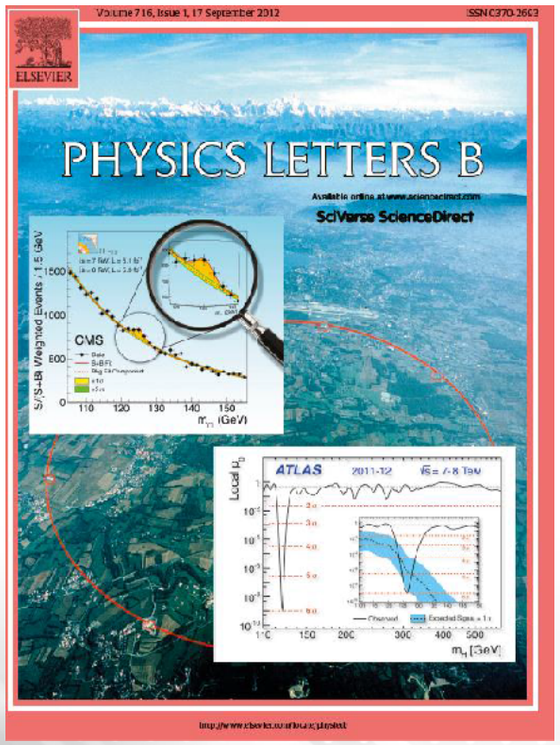
- So far in the journey we read out a raw event - ***what happens next ?***
  - ***Calibration***
  - ***Reconstruction***
  - Searching for “my” ***physics signal***
    - Extracting ***observables of interest***, with some *precision*
  - ***Publication*** of the final statistical analysis
    - Nobel prize (YMMV)

Data's journey

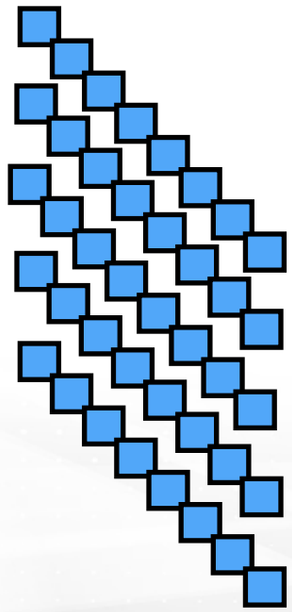




Data Preparation



Distributed computing



U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

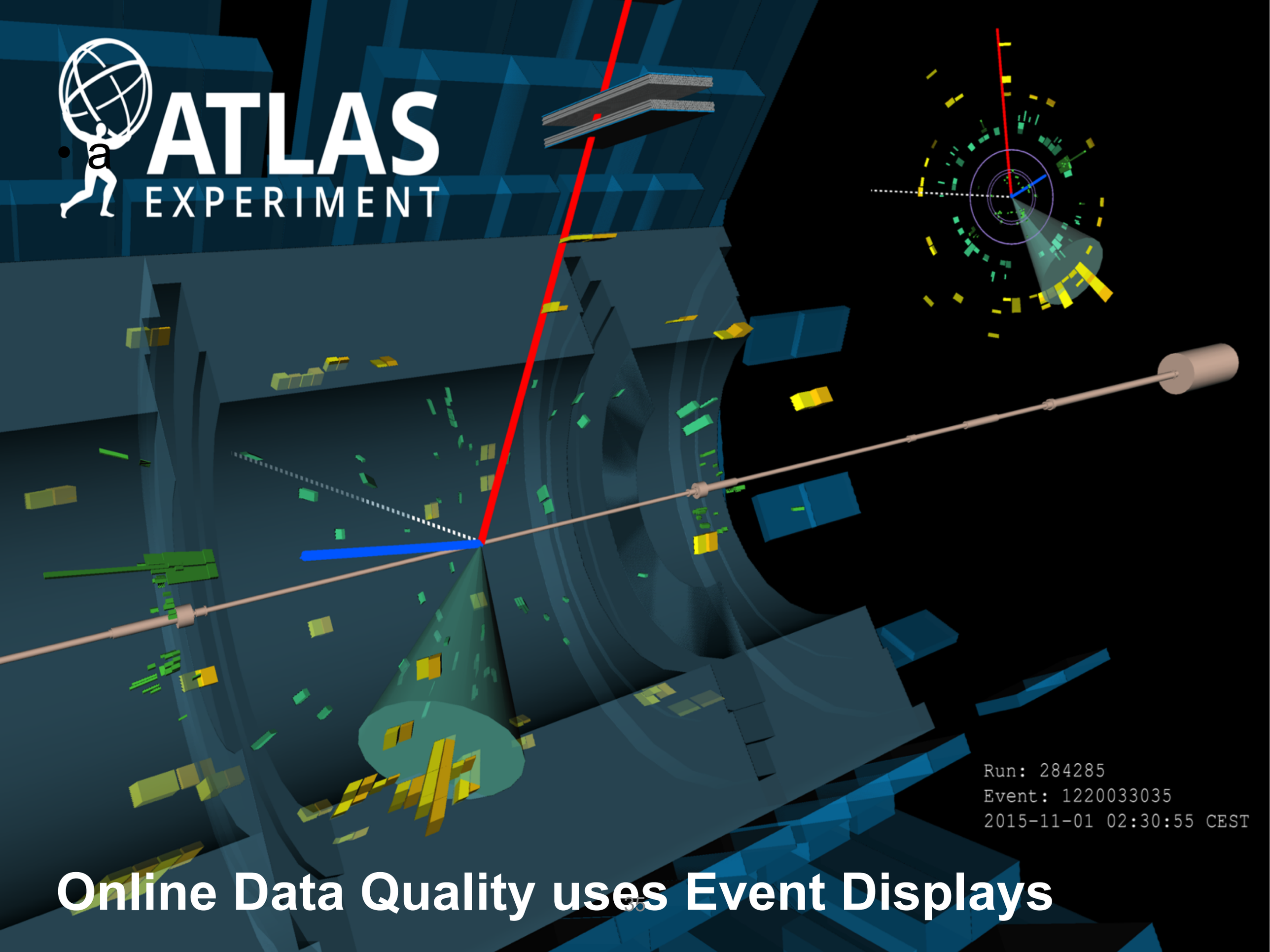
Data Preparation

- Three major steps to **prepare data for physics analysis** and achieve
  - reliable, high quality data (yes, we ***reject*** low quality data)
  - the ***best performance*** from our detectors
  - readiness for ***physics analysis***

1. Make sure that the **data quality** is excellent, also in real time
   - Maximise the amount of data that is useful

2. **Calibrate** the detectors
   - Correct for imperfections in the detectors, account for changes over time, etc.

3. **Reconstruct physics signals** from the data
   - Produce analysis object data which contains physics analysis level information like how many muons does the event have



Online Data Quality uses Event Displays

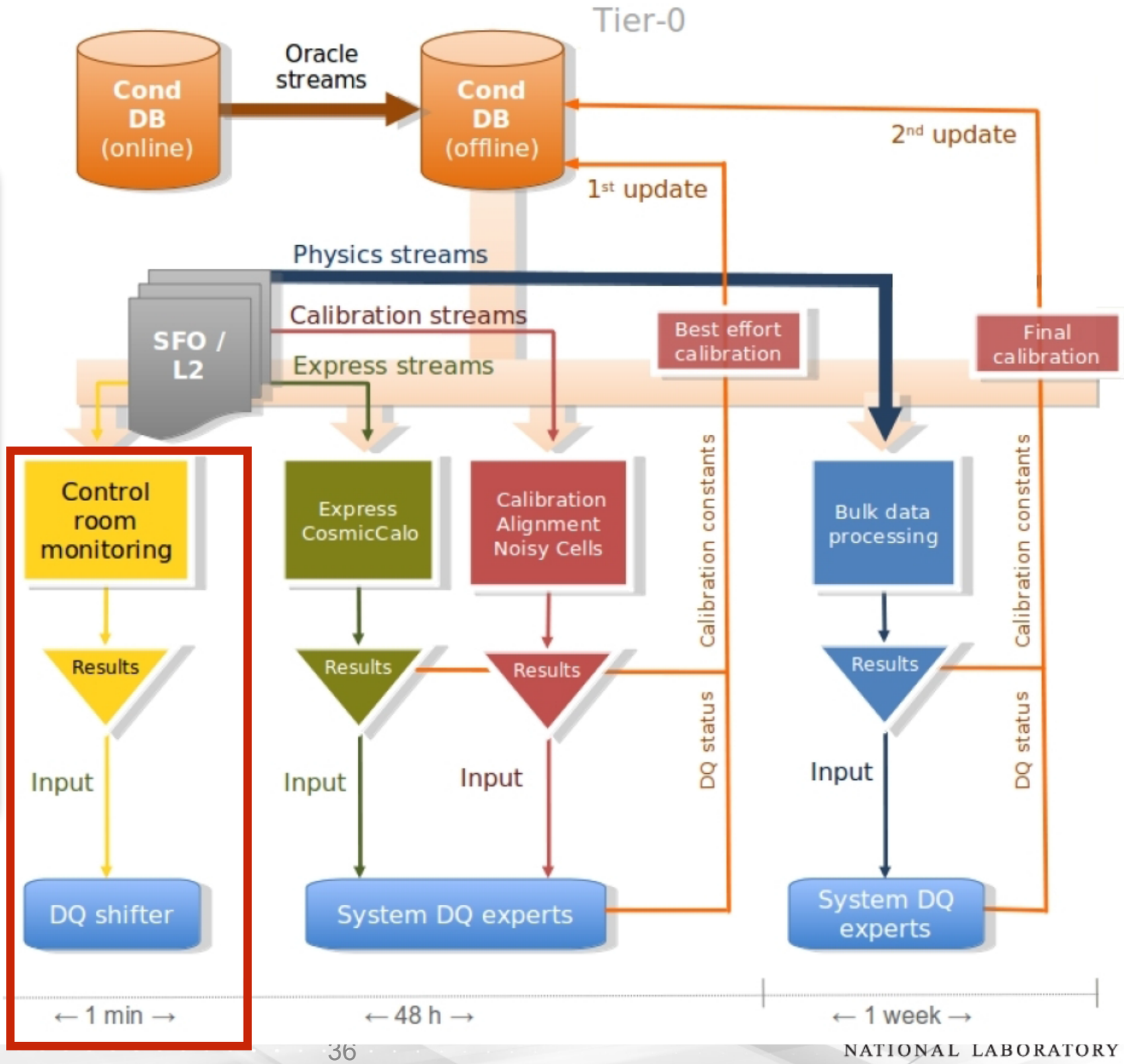
Data Quality in the Control Room

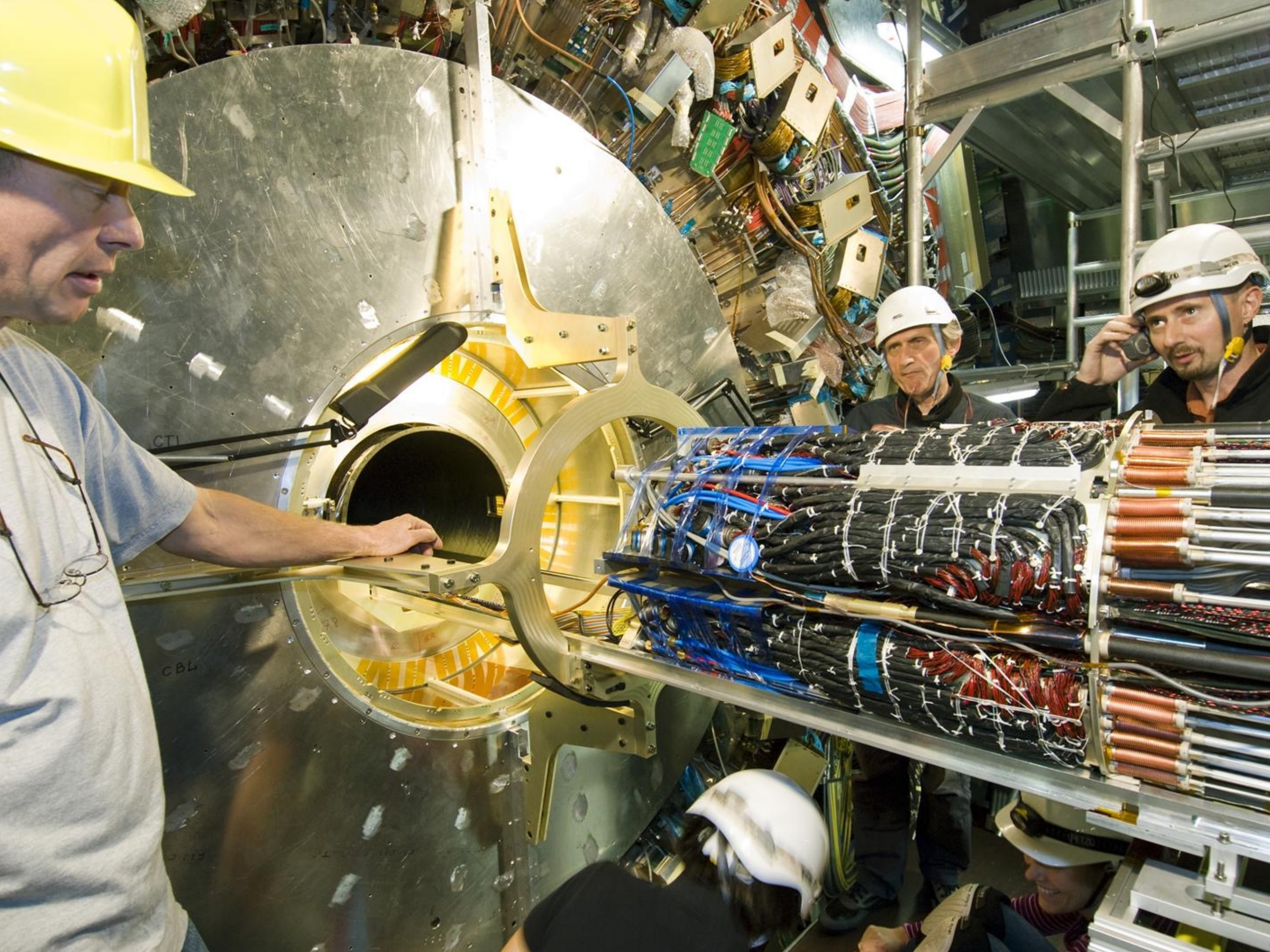
Runs on dedicated computers in the ATLAS Control Room

Runs dedicated algorithms to check data quality

Dedicated data quality expert in the ATLAS Control Room to raise the alarm

Also looks at a small fraction of events using Event Displays to spot problems



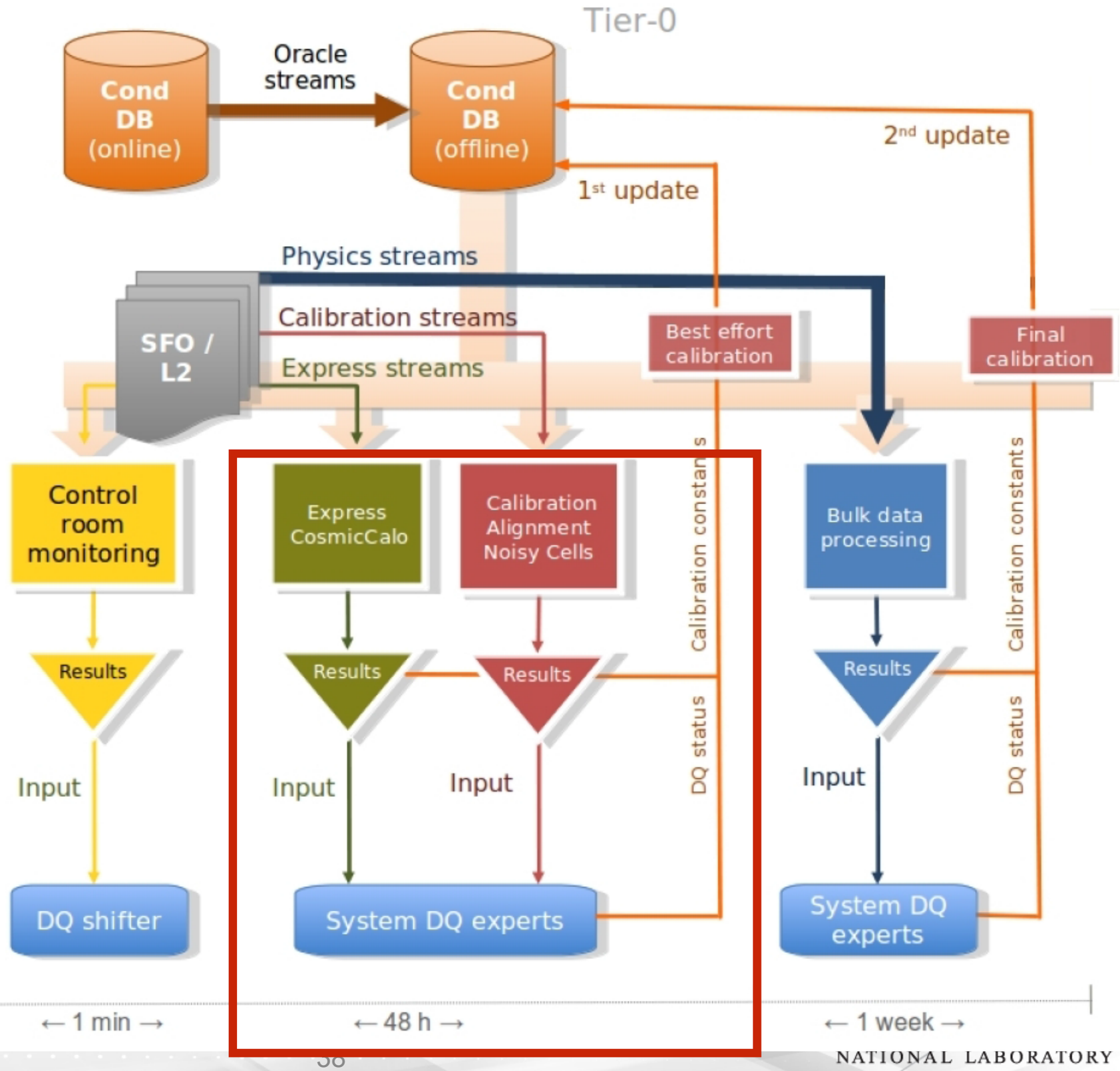


Detector Calibration and more data quality checks

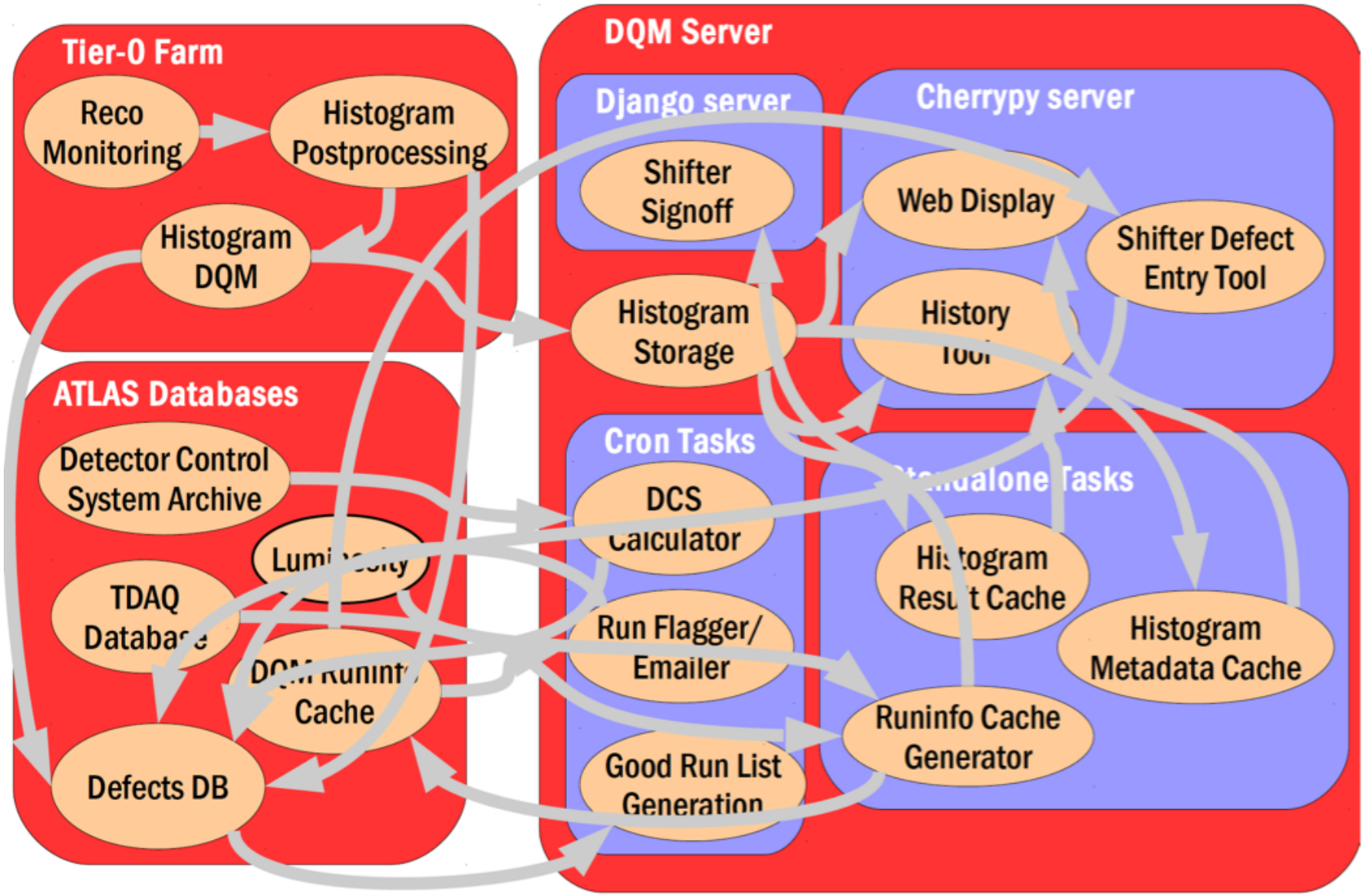
Runs on dedicated computing resources at CERN

Extract detector calibrations

Also assess data quality on more data, ATLAS ~2% data



Data describing data - metadata and databases



- In addition to the **calibration databases**, we need **databases** for several other important metadata tasks, **data quality** is a particularly important example
- Understanding our data requires us to keep precise track of our **metadata** too

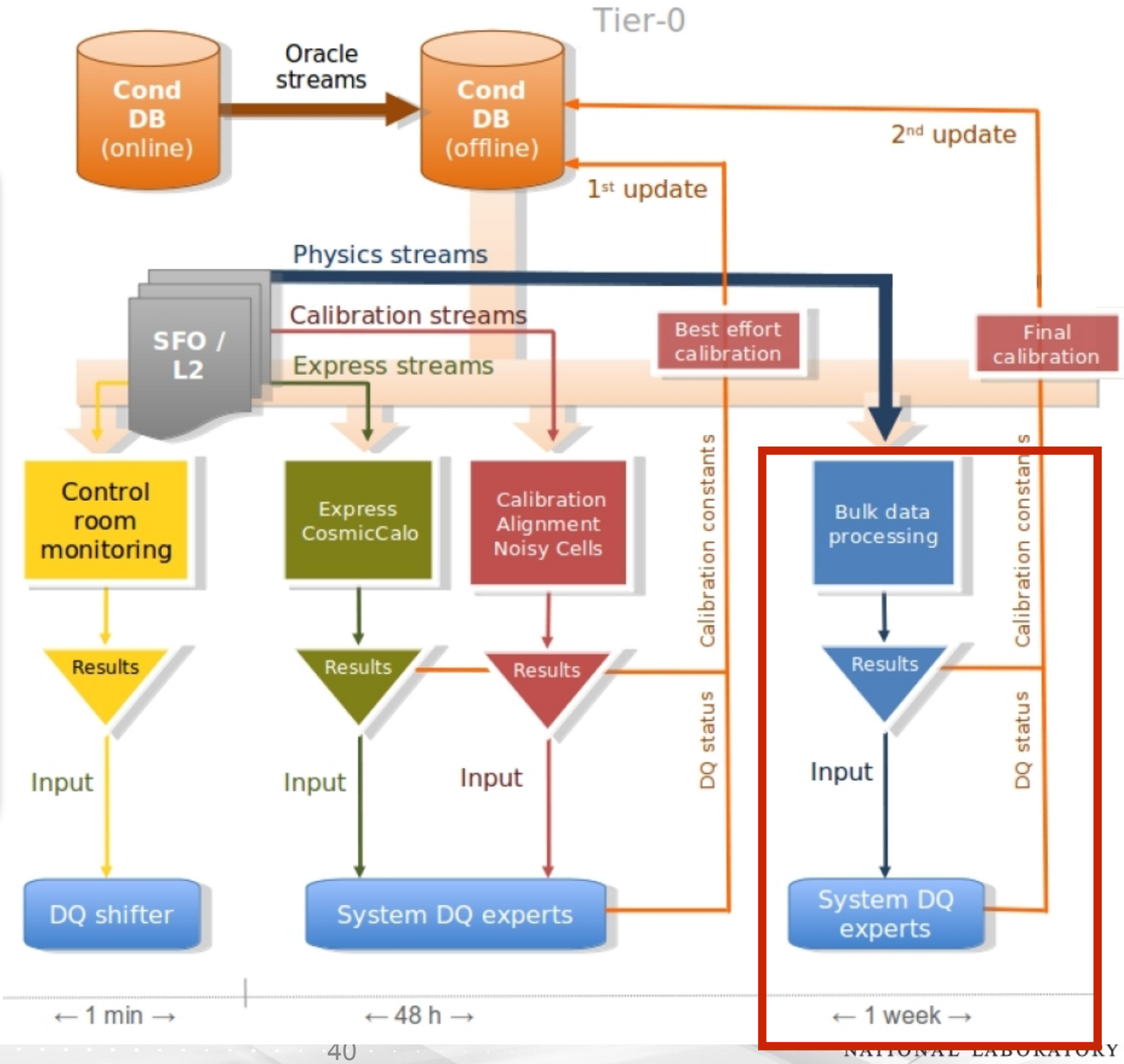
Data Reconstruction and final data quality

Runs on dedicated computing resources at CERN

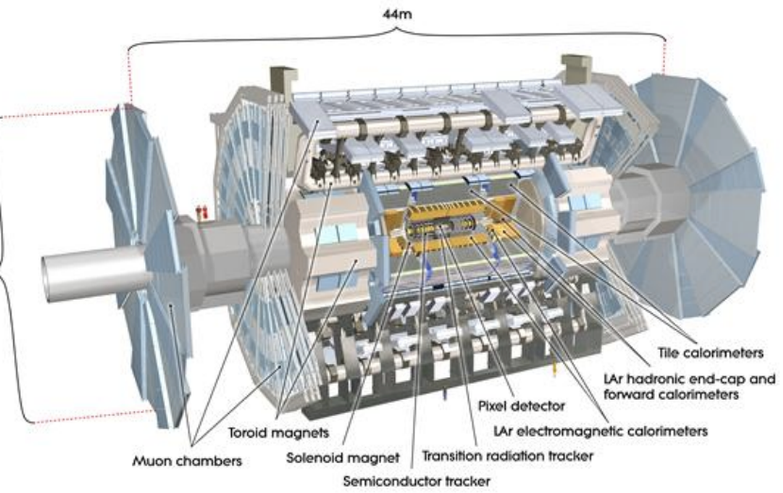
Runs reconstruction algorithms once calibrations are ready

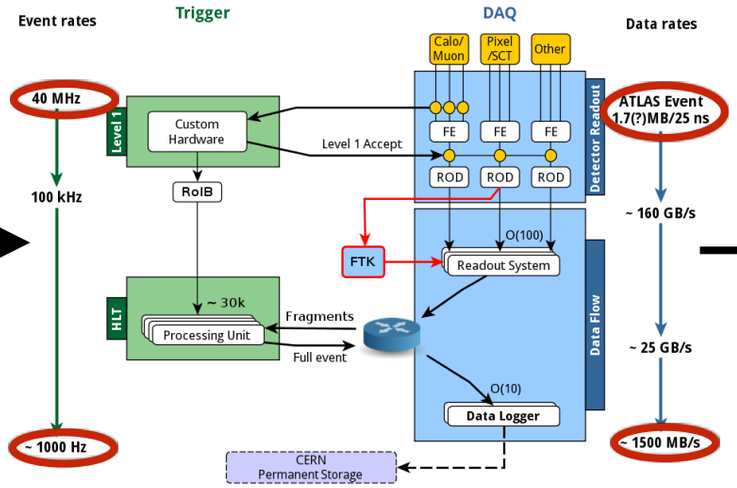
Final data quality assessment made on the results

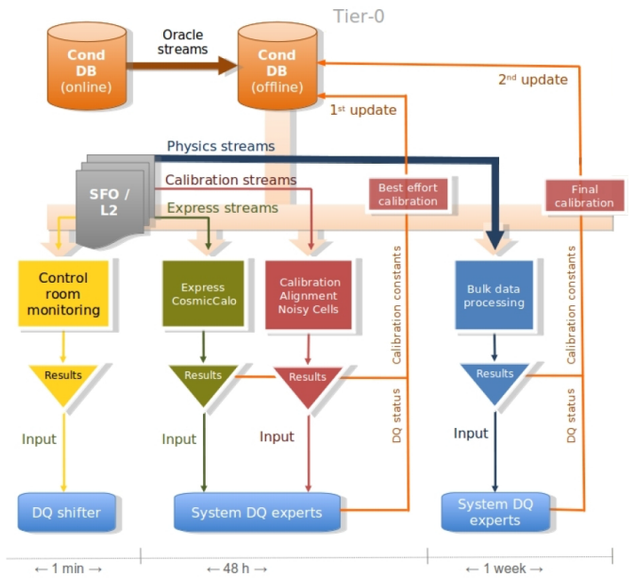
ATLAS typically processes
~60M events per day
~60 TB per day

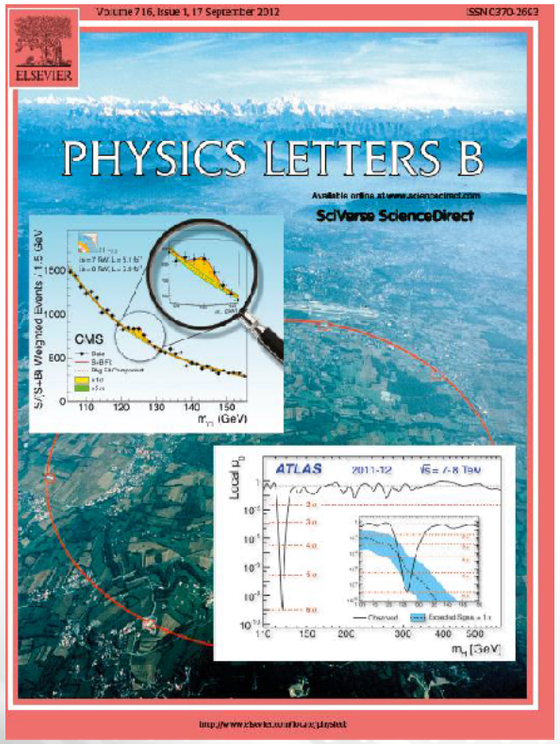


Data's journey

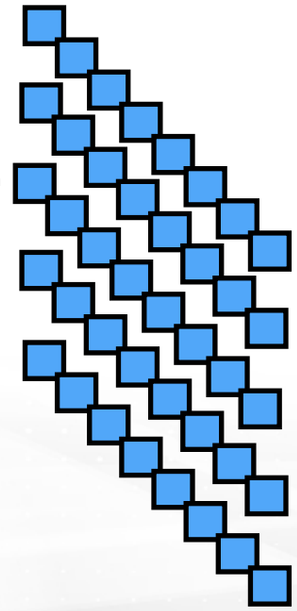








Distributed computing



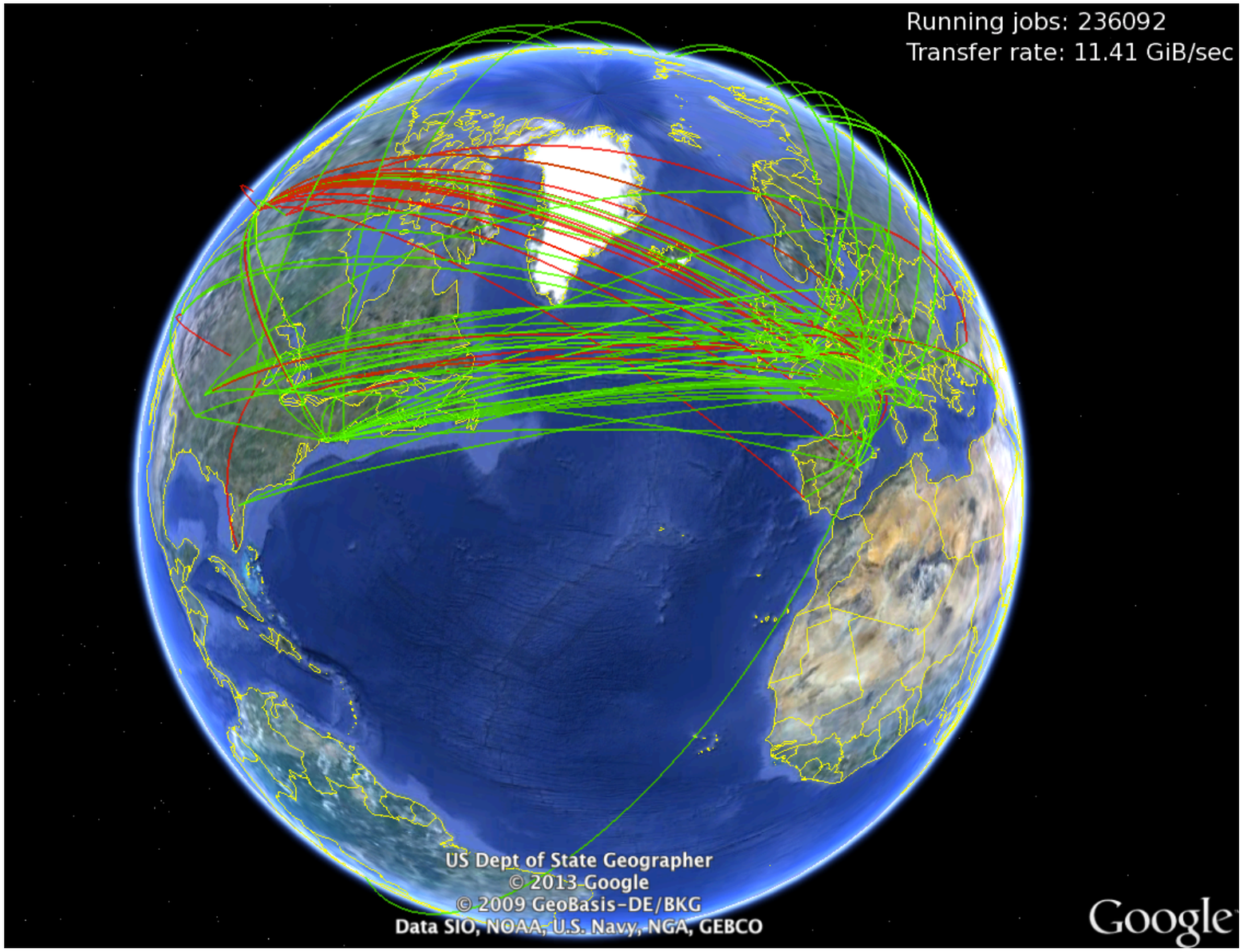
U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

The Worldwide LHC Computing Grid

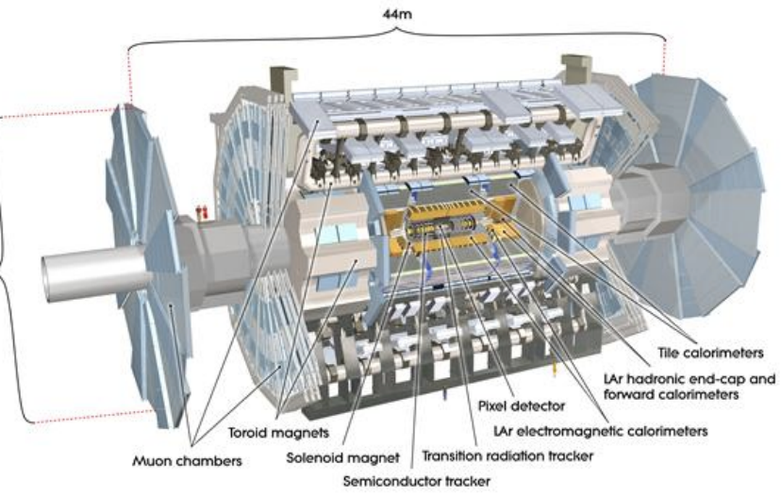
- Now the data has been ***prepared for physics analysis***, it's time to extract our favourite physics signal!

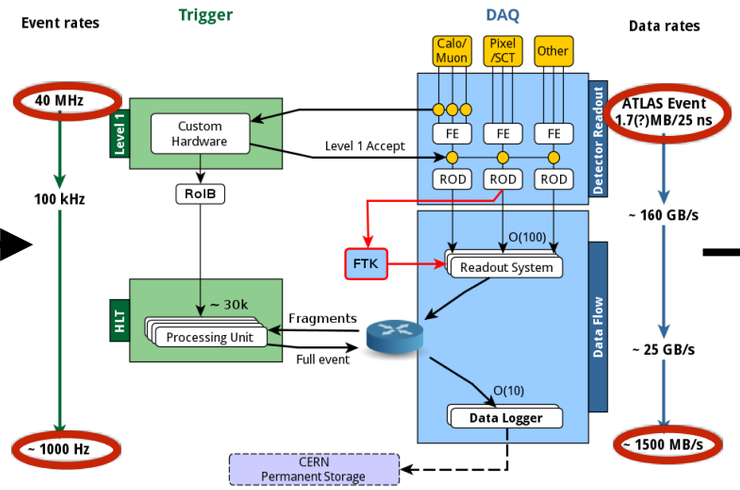
- Many experiments, particularly those at the **LHC**, use computing sites all over the world via **the grid** to
  - harness all of that ***computing power***
  - enable collaborators ***worldwide*** to access the data

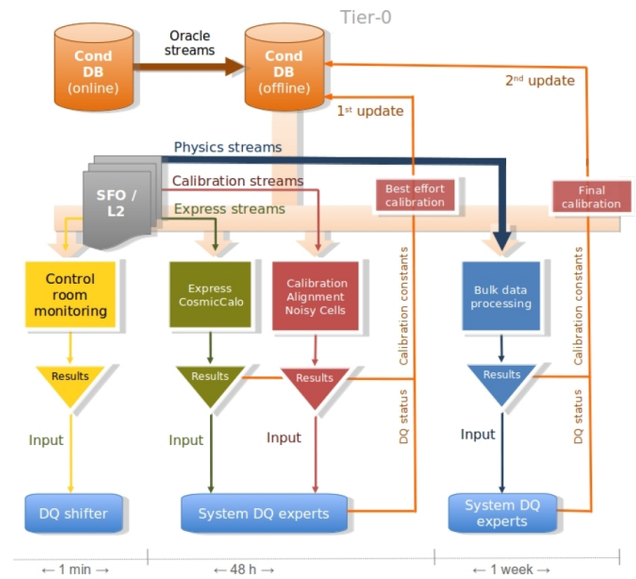


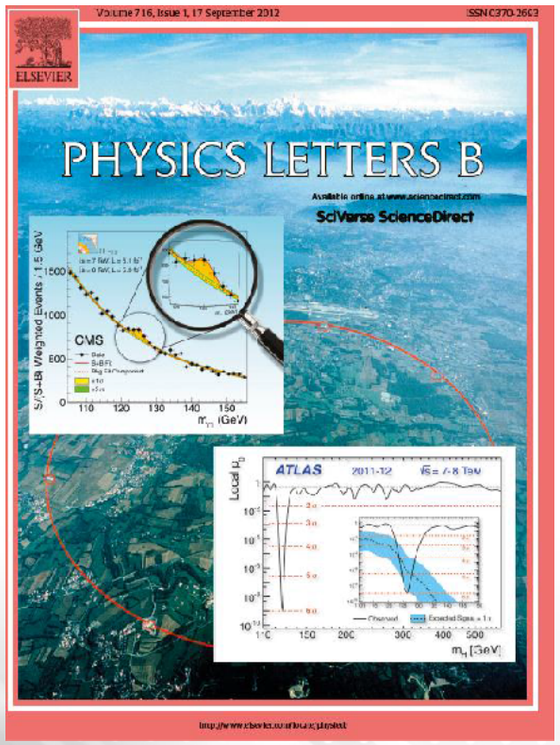
- Image taken from the **WLCG** GoogleEarth Dashboard
  - http://wlcg.web.cern.ch/wlcg-google-earth-dashboard

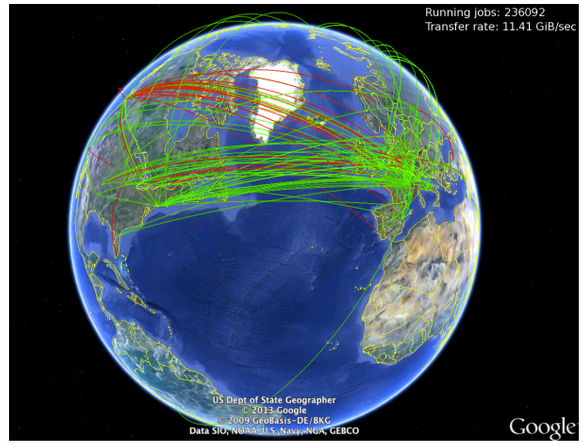
Data's journey









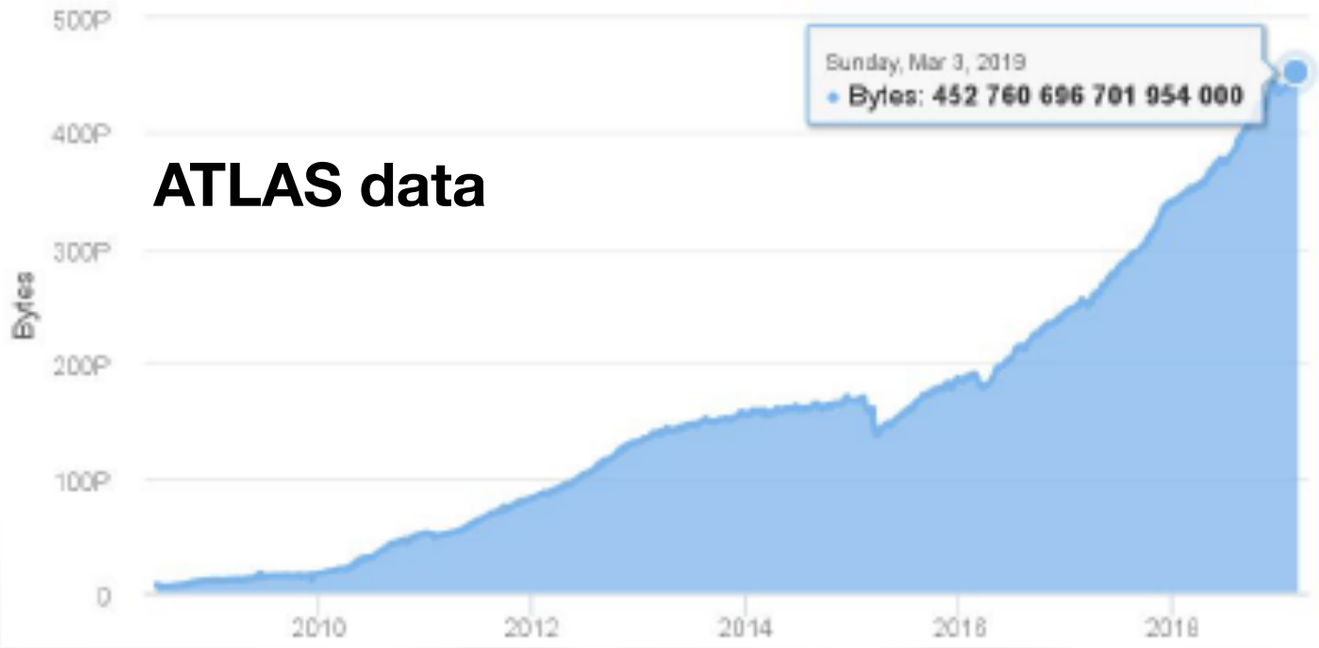


We did it !

- Our data is calibrated and with good data quality
- and we've reconstructed the physics objects in the data
  - ***it is reliable, accurate and ready for physics analysis***
- ***More detail on these topics in Lecture 2***

- ***Now we can extract our measurements in Lecture 3***

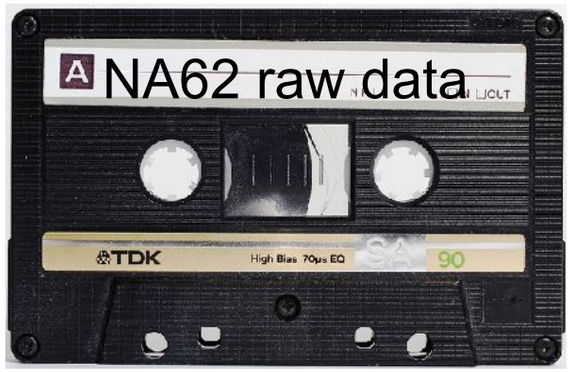
- ***Question****: How long would it take to read all of the ATLAS data? (Assume for simplicity you have off-the-shelf SSDs with read speed ~500MB/s)*

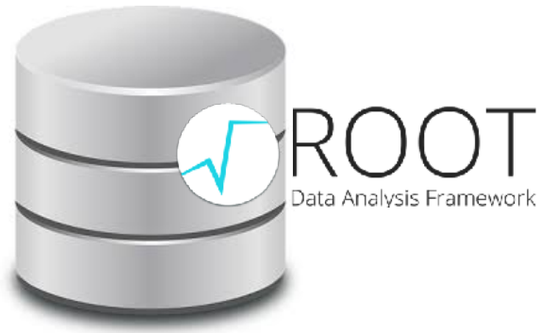


Extra data-centric problem

The NA62 data challenge

$K^+ \rightarrow ?$





events / day (billions)	raw data / day (TB / day)	reco data /day (TB / day)
~1	*?*	*?*

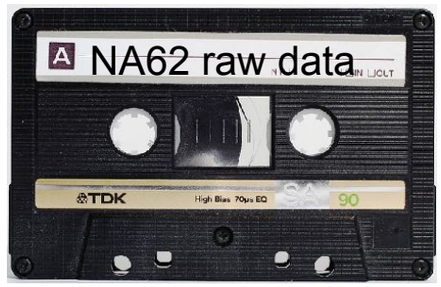
- Unit of data-taking is a CERN SPS burst (aka spill), which lasts ~3.5 seconds
- Take 4000 SPS bursts / day, 250k events per burst
- Event size is 10kB

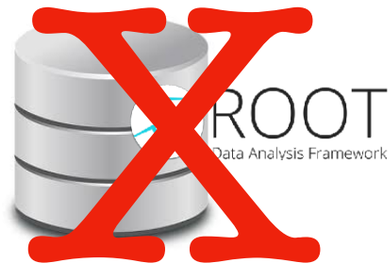
- One DAQ Run is ~1500 bursts, the granularity at which data is calibrated
  - high precision / high rate detectors are also calibrated at burst level

- Fully reconstructed data in NA62 is ~twice the size of the raw data event size

Questions

$K^+ \rightarrow ?$





events / day (billions)	raw data / day (TB / day)	reco data / day (TB / day)	filtered reco / day (TB / day)
~1	*?*	*?*	*?*

- Reconstructed data is filtered to reduce data volume per filter by ~20
  - 10 filters are written
  - 200 physicists on NA62 perform physics analysis on filtered datasets

- **Calculate** the total amount of filtered data in one year (assume 120 days of running)
- **Calculate** the total time required for all 200 physicists to read their filtered data 20 times (assume they only use one filter each)